## DOCTOR OF PHILOSOPHY

# Computational prediction of human protein-protein interactions

Eckenrode Sokolowski, Tara

*Award date:*
2013

DOCTOR OF PHILOSOPHY

# Computational prediction of human protein-protein interactions

Tara Eckenrode Sokolowski

2013

University of Dundee

# COMPUTATIONAL PREDICTION OF HUMAN PROTEIN-PROTEIN INTERACTIONS

By

Tara Eckenrode Sokolowski

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

AT

UNIVERSITY OF DUNDEE

DUNDEE, UNITED KINGDOM

FEBRUARY 2013

# Contents

# List of Tables

or a very low transitive score and a cluster score that increases in value as the number of bin increases. Likewise, bins in row 1 (light pink) contains pairs with no or a very low cluster score and a transitive score that increases in value as the bin number increases. Ideally, the proportion of positives:negatives in the 'no transitive' bins (light blue) and the 'no cluster' bins (light pink) should show a higher number of positives to a low number of negatives, indicating that the method used to group pairs in that bin is able to discriminate between scoring positive and negative examples.

# List of Figures

negative (yellow) examples assigned during training of the neural network with the SCG learning method with 50 hidden nodes is significantly different (KS-test: D = 0.417, p-value = 2.2 e-16).

**3.16 Histogram of distribution of scores for predictions in the full blind test set.** The distribution of positive (pink) and negative (yellow) scores as calculated by the network trained on the EqualFiltered dataset with the SCG learning method are shown. 135

**3.17 Accuracy vs. Precision plot for SCG output scores in the full blind test set.** The accuracy (left-hand y-axis and red line) of predictions made on the full blind test set (6523 positives and 6523 negatives) was calculated by dividing the sum of true positives and true negatives by the total number of positives and negatives for output scores between 0.0 and 1.0 (x-axis). The precision (right-hand y-axis and blue line) of predictions was calculated by dividing the number of true positives by the total number of true positives and false positives at output scores between 0.0 and 1.0. As the cut-off threshold increases, the accuracy increases such that positive and negative pairs are predicted correctly. Conversely, the precision decreases such that at higher cut-off thresholds, fewer positives are predicted overall. 136

**3.18 ROC plot comparing predictions in the EqualFiltered blind test for one-versus two-stage predictors.** ROC curves for the two methods of incorporating network analysis into the neural network PIPs framework with the transitive scores alone predictions for the 0.5 Network (pink, AUC=0.535) and 0.7 Network (orange, AUC=0.589) and with the second neural network step for the 0.5 Network (purple, AUC=0.748) and 0.7 Network (cyan, AUC=0.668) assessing accuracy of predictions in the EqualFiltered blind test set. As a comparison, the ROC curves for the outcomes of the three previous predictors trained on the EqualFiltered dataset with the SCG (red, AUC = 0.775), BackpropChunk (green, AUC = 0.739) and Std_Backpropagation (blue, AUC = 0.780) learning methods on the same blind test set are also shown. While an unpaired T-test comparison of the SCG and Transitive method ROC curves indicates no significant difference (D = 1.814, dof = 3988.667, p-value = 0.070), the lower ROC profiles and AUC value for the Transitive method suggests it predicts less accurately than the one-stage neural network. Curves and calculations were computed with the R package pROC (Robin *et al.*, 2011). 140

**3.19 Comparison of distributions scores assigned to positive and negative training pairs during training the EqualFiltered SCG method with and without the Transitive NN.** The distribution of scores assigned to the positive (pink) and negative (yellow) pairs in the EqualFiltered training dataset during training without the Transitive analysis component (left), with the 0.5 Network and Transitive analysis component (right) are plotted above. While both the SCG learning method without the Transitive analysis and the 0.5 141

Network with the Transitive analysis appear to have assigned the majority of positives with high scores and negatives with low scores, there is a significant difference between both the positive (Wilcoxon T-Test p-value = $1.38e^{-14}$, W=13611098) and negative (p-value = $2.2e^{-16}$, W=10809970) score distributions between the methods.

# Acknowledgements

The work previously completed on PIPs was conceived and started by Dr. Michelle S. Scott and Professor Geoffrey J. Barton and continued by Dr. Mark McDowall.

Thank you to Dr. Tom Walsh for his patience with my questions and emails and for keeping me on track with the technical side. Additionally, thank you to Dr. Chris Cole for his introduction into the world of neural networks. Thank you to Dr. Nick Schurch for his help assorted bits and pieces and suggestions whenever it was needed. Thank you to Dr. Jim Procter for his help with Java and the odd questions here and there. Thank you to the whole Barton group for your advice, questions, answers and suggestions – at some point, each of you have offered some piece of insight (or a large chunk of insight) that has helped me with the science or technically. Thank you to Geoff, both for supervising this work and for putting up with me and going out of his way for me.

Mostly though, the main thanks I have to offer is for the continued support of the entire group throughout these few years. Coming into something I knew almost nothing about, terrified, a bit shell-shocked and probably over my head (or so I thought), each member of the group made an effort to ease the transition and allow me to find my feet. During the times where I have felt that like the duck on the water, calm on surface and yet treading my little feet so rapidly to stay afloat, it has been the easy questions and concerns that have kept me going. While I feel that I sometimes look at Scotland and Dundee and my mind switches to this black-and-white reel where everything is dark and grey, what I will look back at and remember is not the rain or cold or endless days and

nights, but the times in between that have made it all worthwhile. Thank you to Nick and Rhoda, who have somehow known exactly when I need the innocence of a child or a trip out to make me remember that it's not always so serious, and that there is a huge world outside. Thank you to Jim, who has seemed to always know what to ask and when. Thank you to Nancy, for not only bringing to light that I had US tax returns to file, but also for listening to me for hours on end. Thank to you Jaleh, for being there when I needed anything from a Disney movie, a dress, coffee or someone who just understood. Thank you to Kim, who has listened to me on numerous occasions and been the support I have needed.

Finally, thank you to my family. To Richard, who has been there day in and day out when I am happy, sad, annoyed, frustrated, tired or cranky, with a smile on his face and some childlike excitement to offer that perfect balance that keep me goin. To my father, who has spent hours of his life as a patient listener, sounding board and compass, on Tuesday nights, walks with the puppies, in the leather chair and any and everywhere in between. To my mother, who has always cared and been there in any way she could. And to Steve, who makes me the luckiest sister in the world to have a brother who not only sends me pictures of baby animals when he knows I'm in an afternoon slump, but also watches out for me in a quietly protective way whether I am in the same room or across the sea.

UNIVERSITY OF DUNDEE

COLLEGE OF LIFE SCIENCES

I certify that Tara Eckenrode Sokolowski has satisfied all the terms and conditions of the relevant Ordinance and Regulations to qualify in submitting this thesis in application for the degree of Doctor of Philosophy.

Dated: February 2013

Research Supervisor: _____

Prof Geoffrey J. Barton

UNIVERSITY OF DUNDEE

Date: February 2013

Author: Tara Eckenrode Sokolowski

Title: Computational Prediction of Human Protein-Protein Interactions

Department: College of Life Sciences

Degree: Ph.D. Bioinformatics

I hereby declare that the work described in this thesis is my own; that I am the author of this thesis; that it has not previously been put forward in submission for any other degree or qualification; and that I have consulted references herein.

Tara Eckenrode Sokolowski

# Abstract

Over the past decade, knowledge of the human genome has grown exponentially. While identifying individual genes and their protein products is crucial, understanding how these entities exist within the context of other molecules within the cell provides valuable insight into their functional significance. In particular, mapping the intricate web of interactions between proteins (or the 'interactome'), allows for an understanding the roles of individual proteins within specific cellular processes and the potentially negative implications when these processes cannot occur. At the present time, approximately 40,000 binary, protein-protein interactions have been identified in human through low- and high-throughput, lab-based experiments; however, this number represents only a fraction of the estimated 600,000 protein-protein interactions thought to occur. With the high number of potential protein-protein pairing, experimentally testing each possible interaction is a time-consuming and near-impossible task. As a result, several computational methods have been developed to predict probable interactions for experimental verification.

Previously, our group developed PIPs, a predictor of protein-protein interactions in human based on a naive Bayesian framework that has undergone two version releases (Scott *et al*., 2007, McDowall, 2011). In this thesis, a third version of PIPs, PIPs v. 3.0, is described. In addition to an update of the included data, PIPs v. 3.0 contains a new network analysis component, the TransMCL (Z) module, that combines the previously separate Transitive module (and associated EOCT predictor) introduced in version 1.0 and Cluster module (and associated EOCM predictor) introduced in version 2.0. This new module has allowed the two previously separate PIPs predictors to be merged into

one method (the EOCZ predictor). In total, the new EOCZ predictor identifies over 500K significant interactions, made up of those predicted by the EOCT and EOCM predictors individually as well as a new set of interactions.

Additionally, this thesis describes the development of PIP'NN, a new protein-protein interaction predictor built on a neural network framework with the data incorporated into PIPs. Overall, PIP'NN performs slightly better than the three PIPs predictors on multiple blind tests of varying sizes. PIP'NN identifies both interactions predicted by the three PIPs methods as well as a set of new interactions. As a result, PIP'NN is able to stand on its own as a new predictor of human protein-protein interactions or in conjunction with PIPs as a method to further narrow down the set of predicted interactions.

Finally, this thesis describes the practical implementation of PIPs and PIP'NN through collaborations with two groups within the University of Dundee that have identified sets of potential interactions of interest for experimental confirmation. While these interactions have yet to be confirmed, both studies offer a proof of concept of how the predictors can be incorporated into lab-based interaction identification protocols. Additionally, the new PIPs web server will allow outside groups access to the updated PIPs prediction database.

Overall, the work described in this thesis has built upon previous work both within and outside of the University of Dundee to further the identification of novel protein-protein interactions in human and increase the understanding of the human interactome.

# Chapter 1

# Introduction

## Preface

This chapter presents an introduction to the field of protein interaction prediction and the scope of this thesis. First, the general principles of protein interaction are discussed, followed by details of lab-based and machine learning methods for interaction detection and prediction, with specific focus on the Bayesian and neural network methods for prediction. Finally, previous development of the PIPs predictor for human protein-protein interactions before the start of this project is described.

# 1. Introduction

## 1.1 Protein-Protein Interactions

Following the release of the first draft sequence of the human genome in 2001 (Lander *et al.*, 2001), enormous progress has been made in gaining a more comprehensive understanding of genes, their proteins products and, most importantly, their functional roles in the cell. However, despite ongoing efforts, this understanding is still far from complete.

One of the major stumbling points to cracking the human genetic code is its complexity when compared to genomes from other species. In particular, the majority of biological work has been centred around working with a range of model organisms, whose size and number of encoded genes vary drastically from human (Table 1.1, below). While the human genome contains approximately 3.3 billion basepairs, less than 2% corresponds to an estimated 21,000 protein-coding genes encoded in short exons interspersed within long, non-coding segments (Lander *et al.*, 2001, Flicek *et al.*, 2012). Unlike the genomes of other model organisms, for example *Saccharomyce cerevisiae* (yeast) and *Escherichia coli*, which contain long open reading frames and low or no introns respectively, the high signal-to-noise ratio of coding to noncoding genes has made accurate annotation of the human genome a difficult task. As a result, multiple ongoing efforts exist to dynamically revise the genome by either manual re-annotation (e.g. HAVANA (Wilming *et al.*, 2008)) or through an automated pipeline (e.g. Ensembl (Flicek *et al.*, 2012)). In 2007, an expansion of the ENCODE project (ENCODE Project Consortium *et al.*, 2007), GENCODE, was started to annotate the entire human

genome with high accuracy by merging manual curation, automatic annotation and experimental data into one process (Harrow *et al.*, 2012).

| Species | Genome Size (Base Pairs) | Coding Genes | Non-Coding Genes | Assembly |
|---|---|---|---|---|
| *Homo sapiens* (Human) | 3,300,551,249 | 20,476 | 22,170 | GRCh37.p8 (Feb 2009) |
| *Mus musculus* (Mouse) | 3,478,998,185 | 23,153 | 8,662 | GRCm38 (Jan 2012) |
| *Dania rerio* (Zebrafish) | 1,505,581,940 | 26,163 | 6,041 | Zv9 (Apr 2012) |
| *Rattus norvegius* (Rat) | 2,507,066,667 | 22,938 | 4,828 | RGSC 3.4 (Dec 2004) |
| *Gallus gallus* (Chicken) | 1,050,947,331 | 16,736 | 1,102 | WASHUC2 (May 2006) |
| *Xenopus tropicalis* (Frog) | 1,358,329,334 | 18,429 | 1,282 | JGI 4.2 (Nov 2009) |
| *Drosophila melanogaster* (Fruitfly) | 168,736,537 | 13,917 | 1,141 | BDGP 5 (Apr 2006) |
| *Saccharomyces cerevisiae* (Yeast) | 12,157,105 | 6,692 | 413 | EF 4 (Sept 2011) |
| *Caenorhabditis elegans* (Worm) | 103,021,950 | 20,517 | 23,871 | WBcel215 (Oct 2010) |
| *Escherichia coli K12* (Bacteria) | 4,738,834 | 4,258 | N/A | ASM584v1 (Oct 2011) |
| *Arabidopsis thaliana* (Plant) | 135,670,229 | 27,416 | 1,359 | TAIR10 (Sept 2010) |

**Table 1.1: Comparison of genome compositions between model organisms.** Details of the number of base pairs (column 2), coding (column 3) and non-coding genes (column 4) in the genomes of selected model organisms (column 1) are provided according to their most recent release (column 5). All information was extracted from the Ensembl resource and is current as of November 2012 (Flicek *et al.*, 2012).

With these ongoing efforts, a growing number of proteins in human have been identified. While the primary resource for human protein annotation, the International

Protein Index (IPI) (Kersey *et al.*, 2004) was established with the release of the initial genome sequence, it has been recently deprecated in favour of the Universal Protein Reference Knowledge Base (UniProtKB) as the central protein resource (UniProt Consortium, 2012). UniProtKB is split into two sets: UniProtKB/TrEMBL, an unreviewed and computer-annotated collection of protein sequences translated from coding sequences included in either the EMBL-Bank, GenBank or DDBJ databases, and UniProtKB/Swiss-Prot, a reviewed, non-redundant collection of manually annotated proteins (UniProt Consortium, 2012). As of early October 2012, UniProtKB/Swiss-Prot and UniProtKB/TrEMBL contain 20,235 and 110,812 human protein sequences, respectively (http://web.expasy.org/docs/relnotes/relstat.html).

While annotating genes and identifying their protein products is a necessary initial step, acknowledging that a protein exists is only one piece to understanding its functional significance. Although attempts have been made to assign functions to proteins based on features alone (Pavlidis *et al.*, 2002), additional, valuable insight can be gained from considering how the protein behaves within the context of its environment and how it physically interacts with other proteins or molecules. Both transient, or short-term, and stable, or longer-term, physical interactions underlie all biological processes and are critical to maintaining the dynamic nature of the cell. Therefore, it is necessary that as best of an understanding as possible of the full network of physical interactions between all proteins, referred to as the 'interactome', is achieved. While the human interaction network is estimated to be comprised of about 600,000 unique, binary interactions (Stumpf *et al.*, 2008), only approximately 40,000 of these have been confirmed experimentally, leaving a large number of potential interactions yet undiscovered.

Assembling a comprehensive interactome for a species is dependent upon both how much is known about each protein individually and how it relates to other proteins. As a result, the majority of previous work has focused on constructing the interactomes for model organisms (i.e. yeast, worm, fly and human) that have the greatest data availability (Kiemer & Cesareni, 2007). Ideally, known interactions in one species should be transferrable to other species. However, a 2006 study by Gandhi et al. comparing the overlap of experimentally identified protein-protein interactions between yeast, worm, and fly to human showed otherwise, with only 42 of the known interactions conserved between human, worm and fly and 16 conserved between human, worm, fly and yeast (Gandhi *et al.*, 2006). While these low numbers could be attributed partially to a lack of orthologues for many of the proteins, it has also been suggested that even low rates of gene mutation and duplication can lead to the addition and loss of interactions that were conserved after the species diverged (Wagner, 2001).

Therefore, while knowledge of protein interactions in one species can suggest potential interactions in another, conservation alone is not enough to construct a complete interactome. Although the size and complexity of the human genome makes building its interactome difficult, knowing how proteins function together has valuable implications for not only understanding large- and small-scale processes within the cell when it functions normally, but also the effects on these processes when interactions cannot occur.

# 1.2 Experiment-Based Protein Interaction Identification Techniques

In order to identify and verify protein-protein interactions, there are several lab-based experimental methods split into two main categories: 1) low-throughput experiments, which aim to identify a specific protein-protein interaction or complex and 2) high-throughput experiments, which aim to identify at a large set of potential interactions at once (Xia *et al.*, 2004). A summary of the main methods of both categories is provided below.

## 1.2.1 Low-Throughput Methods

### 1.2.1.1 Immunoprecipitation

In immunoprecipitation (shown schematically in Figure 1.1, below), a cell lysis sample is incubated with an antibody (yellow) specific to the antigen of a target protein of interest (teal). The sample is then eluted through a column with beads (red circles) coated with an antibody-binding protein (typically the bacterial staphylococcal Protein A or streptococcal Protein G), such that complexes with the bound antibody adhere to the protein on the beads, and unbound molecules are washed away. While immunoprecipitation on its own can identify single proteins of interest, the method can be extended to co-immunoprecipitation (for 'complex-immunoprecipitation', also referred to as 'pull-down assays'), which follow a similar protocol to detect the target protein of interest and any additional non-immunological molecules bound to it at the time (orange). Following both experiments, the eluted proteins and complexes are analysed by SDS-PAGE alone, SDS-PAGE and mass spectrometry or western blotting.

Although co-immunoprecipiation/pull-down assays can successfully identify protein complex interactions in their natural cellular environment, the method is highly dependent upon the experimental reagent and protocol followed and suffers from a high rate of detection of environmental contaminants, non-specifically bound proteins or proteins bound immunologically and not to the target protein. Additionally, protein interactions involving individual proteins in different subcellular localisations and transient or unstable interactions are not able to be purified.

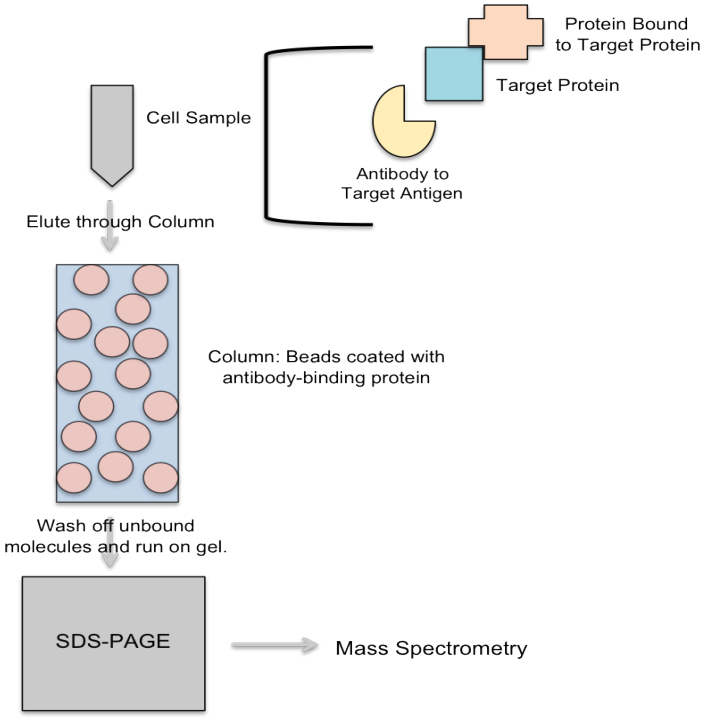


**Figure 1.1: Immunoprecipitation and co-immunoprecipitation.** In immunoprecipitation experiments, a cell sample is incubated with an antibody (yellow) to the antigen of the target protein of interest (teal). The sample is then eluted through a column containing beads (red) coated with an antibody-binding protein. After washing away unbound molecules, the eluted sample is analysed (for example, through SDS-PAGE and mass spectrometry) to identify the proteins retrieved. Co-immunoprecipitation experiments follow a similar protocol to immunoprecipitation experiments except after the sample is eluted, both the target protein of interest and any proteins bound to it are also analysed.

#### 1.2.1.2 Structural Visualisation of Protein Complexes

Structural visualisation is the most reliable method of verifying a protein complex. Three main methods exist: x-ray crystallography, cryo-electromicroscopy and nuclear magnetic resonance.

In x-ray crystallography, a target protein or protein complex of interest is purified to a high concentration and then grown as crystals (Smyth & Martin, 2000). When an x-ray beam is shone on the crystal, the light diffracts, and the resulting patterns are processed to determine the symmetry of the crystal and a map of the electron density. From the resulting electron density map, a three-dimensional structure of the protein or protein complex is built and refined. While x-ray crystallography allows determination of protein structures to resolutions as low as 1.5 Å, it is limited by if the protein of interest is able to be purified and crystallised or not (Smyth & Martin, 2000). Currently, the Protein Data Bank (PDB), a centralised data store for protein structures, contains 3562 human protein structures (when filtered for 90% sequence similarity) determined by x-ray crystallography (as of November 2012) (Berman *et al.*, 2000).

In cryo-electromicroscopy (cryo-EM) experiments, a sample containing a target protein or complex of interest is placed in liquid ethane at the temperature of liquid nitrogen so that is it suspended in its native molecular state (Frank, 2002). The embedded sample is then fired with an electron beam, which moves through the empty areas and gives a two-dimensional image of the protein or protein complex. By tilting the sample at different angles, a range of EM images are generated that can be combined to give a three-dimensional structure (Spahn & Penczek, 2009). While cryo-EM structures are typically lower resolution than x-ray crystallography structure, the technique has a

distinct advantage in being able to analyse large macromolecular complexes in their native orientation without the constraints imposed by the crystallisation process (Russell *et al.*, 2004). Currently, there are 12 human protein structures (when filtered to 90% similarity) in the PDB (as of November 2012) (Berman *et al.*, 2000).

Finally, nuclear magnetic resonance (NMR) offers an additional method of structure determination. NMR is based on the Nuclear Overhauser Effect (NOE), where the hydrogen atoms of two nuclei experience an observable magnetic dipole-dipole interaction when they are in close contact, that when considered across a sample can build a three-dimensional structure (Vinogradova & Qin, 2012). In one of strategy used for protein interaction detection, the nitrogen and carbon atoms of one protein are labelled with 15N and 13C while the other is unlabelled, such that atoms can be mapped specifically to the appropriate protein (O'Connell, Gamsjaeger & Mackay, 2009). The current PDB contains 2266 protein structures solved by NMR (when filtered for 90% redundancy, as of November 2012) (Berman *et al.*, 2000).

## 1.2.2 High-Throughput Methods

### 1.2.2.1 Yeast Two-Hybrid Screening

The general principle of yeast two-hybrid screening (Y2H) involves two independent halves that are inactive until they are joined together (Uetz, 2002). Figure 1.2, below, shows schematically the experimental process of Y2H screens. Using the Gal4-*HIS3* method as an example, a DNA binding domain (typically a yeast transcriptional activator protein, Gal4, shown in orange) is fused to the N-terminus of the protein of

interest (the 'bait', shown in purple), and the binding protein's activation domain (shown in green) is fused to the C-terminus of the second protein (the 'prey', shown in teal in A and yellow in B). Each new construct is then inserted in a plasmid that is expressed by the yeast cell. The plasmid encoding the bait protein also includes HIS3, a gene encoding imidazolglycerophosphate (IGP) dehydratase, a necessary enzyme for the biosynthesis of the amino acid histidine.

Since the bait protein only contains the DNA binding domain, it remains inactive unless it binds with the transcriptional activation domain fused with the prey protein (shown in A). If the yeast cells are cultured in histidine-free media, they will only be able to grow if the two proteins interact, HIS3 is activated and histidine biosynthesis can occur (shown in B). The cells are plated and positive colonies are then sequenced to identify which bait-prey coupling is responsible for the HIS3 activation (Rajagopala *et al.*, 2012).

Y2H screening typically follows one of two experimental models: library screening and matrix (also called array) screening. With library screening, a set of cDNAs or DNA fragments act as the prey for one bait, the protein of interest. While library screening is widely employed for identifying interactions, its lengthy experimental protocol makes it time-consuming and impractical for large-scale investigations. Additionally, incorrect incorporation of the cDNA and fragment-containing plasmids into yeast open reading frames during transformation can lead to a high number of non-specific or false interactions.

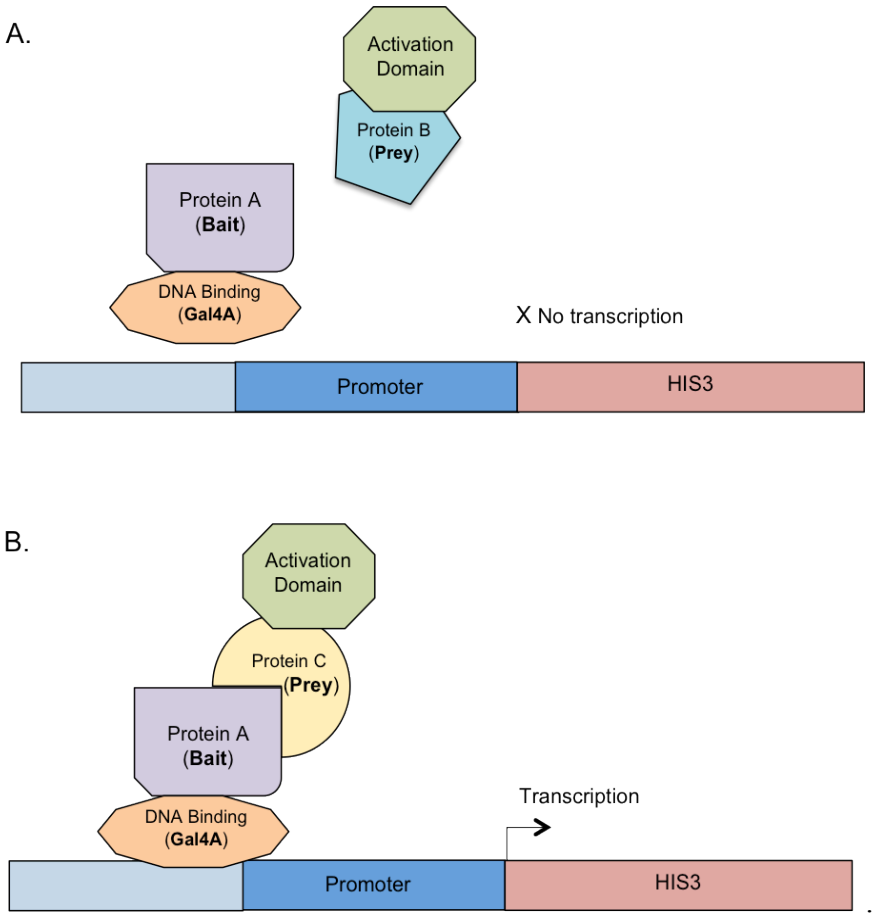


**Figure 1.2: Yeast two-hybrid screening.** In Y2H screening in the Gal4A-*HIS3* system, yeast cells are transfected with a plasmid containing two constructs: a 'bait' and a 'prey' and then grown in Histidine-free media. In order to grow, the cells need to transcribe the *HIS3* gene (red). Transcription of *HIS3* only occurs if the DNA binding domain fused to the bait protein (Protein A in A and B, purple) comes in proximity to the activation domain fused to the prey protein (Protein B in teal in A and Protein C in yellow in B) when the bait and prey interact. Therefore, if the bait and prey interact, *HIS3* is expressed and the cells will grow when plated.

As an alternative Y2H method, array or matrix screening allows a set of cDNAs for prey proteins to be tested systematically for interaction with a bait protein. Rather than co-expressing all prey proteins in the same cells, individual prey proteins are expressed and are plated in a matrix format such that each position is associated with a particular interaction. By knowing ahead of time which interaction corresponds to which positive growth, screens can proceed quicker and be targeted to only proteins suspected to interact (Uetz, 2002).

While Y2H screening is capable of detecting short-lived transient or unstable interactions without the need for large-scale protein expression, the method only allows identification of proteins that can be expressed in the yeast cell nucleus (Mering *et al.*, 2002; Uetz, 2002; Pitre *et al.*, 2008). As a result, the technique falls short with identifying interactions involving transmembrane proteins, proteins expressed in other subcellular locations or proteins undergoing post-translational modifications (Uetz, 2002). Additionally, specific protein families, namely tyrosine kinases, which are toxic at high levels to yeast, are unable to survive with the yeast cellular environment and therefore cannot be tested as bait or prey (Uetz, 2002).

#### 1.2.2.2 Tandem Affinity Purification and Mass Spectrometry

In tandem affinity purification (TAP) and mass spectrometry (MS) experiments, shown in Figure 1.3, the protein of interest is tagged through homologous recombination with two proteins: the calmodulin-binding protein (CBP) (orange) and Protein A (ProtA) (red) from the bacteria Staphylococcus aureus that are separated by the tobacco etch virus (TEV) protease recognition site (yellow) (Rigaut *et al.*, 1999). The cell lysate is then run on a Sepharose column covered with the ProtA-antibody IgG (red circles), so that proteins tagged with the ProtA from S. aureus bind to the column.

The column is then washed to remove any unbound proteins, and the remaining proteins are incubated with TEV protease, an enzyme that cleaves between the two tags on the protein of interest. As a second step to ensure all unbound and contaminant molecules are removed, the lysate is purified again and ran through a calmodulin Sepharose column that binds any proteins with the CBP tag (yellow circles). After washing away any unbound proteins, the lysate is then run on a gel and protein bands are analysed by

MALDI-MS for evidence of interacting protein complexes (Rigaut *et al.*, 1999; Abu-Farha, Elisma & Figeys, 2008).

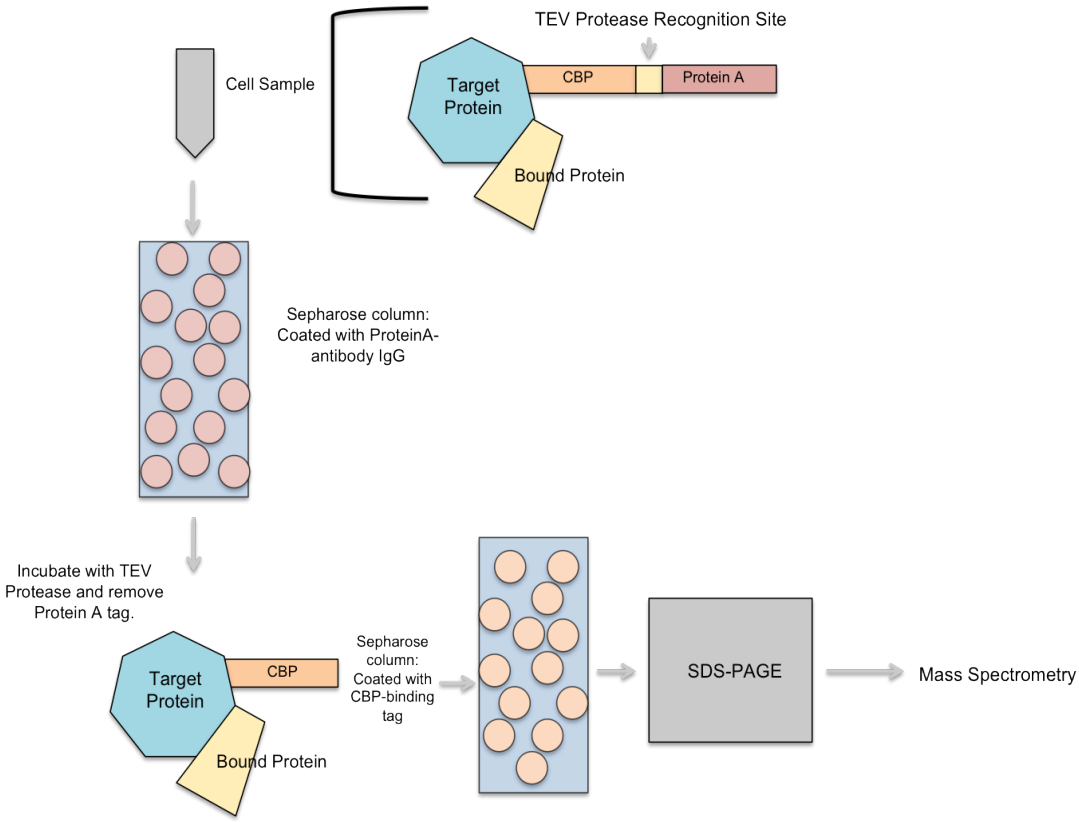


**Figure 1.3: Tandem affinity purification.** In TAP, the target protein of interest is tagged with a construct of two proteins: a calmodulin-binding domain (CBP, orange) and Protein A from S. aureus (red) that are separated by a TEV protease recognition site (yellow), and incubated in a cell lysis sample. Next, the sample is eluted through a Sepharose column coated with an antibody of Protein A, IgG, so that the target protein and any proteins bound to it bind. After washing away unbound molecules, the eluted complexes are incubated with TEV protease, which cleaves off the Protein A domain of the tag. The samples are then eluted through a second Sepharose column coated with CBP-binding proteins to bind the remaining target protein complexes. Samples are then analysed by SDS-PAGE and mass spectrometry to determine which proteins are bound to the target protein of interest.

A second variation on TAP and MS analysis is a one-step procedure in which the proteins of interest are tagged with the FLAG-tag, a short, one kDa peptide sequence that can be recognised by anti-FLAG antibodies bound to a bait protein (Ho *et al.*, 2002). After an immunoprecipitation step, the eluted proteins are run on an SDS-PAGE gel and analysed by ESI LC-MS/MS, another method of mass spectrometry (Ho *et al.*, 2002). While the smaller size of the FLAG-tag reduces the risk that the intrinsic properties of the protein of interest will be altered by the fusion, it also results in a higher number of false positive interactions than TAP with ProtA-CBP tagging (Abu-Farha, Elisma & Figeys, 2008).

While mass spectrometry identification does provide clear indication of interaction and the homologous recombination with TAP has the advantage of proteins expressed under their own promoters, the multiple stages of purification and repeated washes can result in positive interactors being lost in the process (Mering *et al.*, 2002; Abu-Farha, Elisma & Figeys, 2008). Nonetheless, it has proved to a successful technique for characterising large-scale interactions (Gavin *et al.*, 2002; Ho *et al.*, 2002; Ewing *et al.*, 2007).

#### 1.2.2.3 Synthetic Lethality

Synthetic lethality, the principle that if deleting or inactivating one of two non-essential genes does not affect the cell but eliminating both genes is lethal, then the two genes are considered to interact, can be used as an additional, indirect method for identifying protein interactions by way of creating a genetic interaction map of gene functions and pathways (Ooi *et al.*, 2006). Synthetic lethality analyses have been applied across the yeast genome through the synthetic genetic arrays (SGA) and synthetic lethal analysis

by microarray (SLAM) methods (Tong & Boone, 2006; Ooi *et al.*, 2006) and combined with other high-throughput and computational methods (Kelley & Ideker, 2005).

## 1.3 Current Protein Interaction Databases

Human protein-protein interactions that are experimentally verified are included in nine main databases (Bader, Betel & Hogue, 2003; Peri *et al.*, 2003; Chen, Mamidipalli & Huan, 2009; Prasad, n.d.; Croft *et al.*, 2011; Mewes *et al.*, 2011; Szklarczyk *et al.*, 2011; Kerrien *et al.*, 2012; Licata *et al.*, 2012) that are summarised in Table 1.2, below. While all of the resources shown include a set of experimentally validated interactions, each of their methods of curation differs slightly. For example, while the Human Protein Reference Database (HPRD), is manually assembled through literature curation and only includes binary physical interactions, STRING (Szklarczyk *et al.*, 2011) and BOND (formerly BIND) (Bader, Betel & Hogue, 2003) include direct interactions as well as pathway associations, while I2D (Brown & Jurisica, 2007) and STRING also include interactions predicted by their own algorithms (described in more detail below).

A 2009 comparison of BioGRID, MINT, BIND, IntAct, DIP and the HPRD showed that of the resources examined, all but IntAct showed at least two-thirds overlap with the HPRD, the database with the most human protein-protein interactions (Lehne & Schlitt, 2009). Additionally, IntAct and MINT and BioGRID and DIP share over 55% of interactions (Lehne & Schlitt, 2009). Meta-databases, such as the Human Annotated Protein-protein

| Database | Website | Number of Human Protein-Protein Interactions | Information Included |
|---|---|---|---|
| HPRD (Human Protein Reference Database) | http://www.hprd.org/ | 39,141 | Human proteins<br><br>Experimental evidence (co-immunoprecipitation, yeast two-hybrid screening, pull-down assays)<br><br>Protein-protein interactions, post-translational modifications, subcellular localisation, tissue expression, diseases, domains, interactions with nucleic acids and small molecules |
| IntAct | http://www.ebi.ac.uk/intact/ | ~12,000 | Proteins from multiple species<br><br>Includes isoform-specific interactions and protein and non-protein interactions<br><br>Web tools including 'Pay-As-You-Go' algorithm to query for bait suggestions for pull-down assays based on proteins most likely to be inter-connected |
| MINT (Molecular Interactions Database) | http://mint.bio.uniroma2.it/ | 26,666 | Mammalian proteins<br><br>Interactions scored with a confidence score based on number of verifying experiments<br><br>mRNA and promoter interaction, isoform-specific interactions, genetic disease information |
| BIND (Biological Interaction Network Database)<br><br>(now BOND - Biological Object Network Database) | http://bond.unleashedinformatics.com/ | 36,029 | Proteins from multiple species<br><br>Includes binary, complex and pathway interactions<br><br>Attributes of proteins described and marked through ontoglyph symbols from GO terms, functions and the NCBI's Cluster of Orthologous Groups (COG) |
| DIP (Database of Interaction Proteins) | http://dip.doe-mbi.ucla.edu/dip/Main.cgi | 4540 | Interactions for 469 different organisms<br><br>Experimental evidence<br><br>Includes feature to verify experimental and predicted interactions through three techniques (Paralogous Verification Method, Expression Profile Reliability, Domain Pair Verification) |
| MIPS (Mammalian Protein-Protein Database) | http://mips.helmholtz-muenchen.de/proj/ppi/ | 475 | Mammalian proteins<br><br>Experimental evidence - mass spectrometry and yeast two-hybrid interactions not included |

| Database | Website | Number of Human Protein-Protein Interactions | Information Included |
|---|---|---|---|
| Reactome | http://www.reactome.org/ | 5387 | Proteins multiple species<br><br>Only interactions involving enzymes; classifies interactions as 'direct complex', 'indirect complex', 'reactions' and 'neighbouring reactions'<br><br>Ultimately creates a biological pathway network map |
| HAPPI (Human Annotated Protein-Protein Interactions) | http://discern.uits.iu.edu:8340/HAPPI/ | 1,209,463 | Human proteins<br><br>Collation of interactions from HPRD, BIND, MINT, STRING and I2D known and predicted interaction databases<br><br>Interactions assigned a confidence scored based on how many interactions they contain that are part of designated dataset composed of evolutionarily co-expressed pairs of proteins (from MetaGene) |
| APID (Agile Protein Interactions Data Analyzer) | http://bioinfow.dep.usal.es/apid/index.htm | 83,670 | Multiple species<br><br>All known experimentally validated protein-protein interactions from HPRD, BIND, BioGRID, DIP, IntAct and MINT<br><br>Interactive; calculates the weights of the edges connecting proteins by considering connectivity, clustering, GO terms, number of experimental validations and domain-domain interactions |
| STRING | http://string-db.org/ | 3,281,414 | Multiple species<br><br>Includes experimentally validated interactions from MINT, BioGRID, BIND, DIP, HPRD, Reactome, KEGG, IntAct, EcoCyc, NCI-Nature Pathway Interaction Database and GO protein complexes<br><br>Also includes interactions predicted from functional genomics data (i.e. microarrays)<br><br>Computes a confidence score for each interaction link based on how likely the pair are to be in the same KEGG pathway<br><br>Slick interactive interface |

**Table 1.2: Main online databases detailing human protein-protein interactions.** Human interaction counts are current as of late August 2012.

Interactions (HAPPI) database (Chen, Mamidipalli & Huan, 2009) or Agile Protein Interaction Database (APID) (Prieto & Las Rivas, 2006), attempt to address this lack of full coverage by centralising all interactions identified by each of the other databases

into one, aggregated collection. Overall, which database or databases are selected for a specific task should depend upon what is required from the information.

## 1.4 Computational Prediction of Human Protein-Protein Interactions

Overall, the low and high-throughput in vivo experimental techniques described above have been successful in producing large datasets of interactions for *Homo sapiens* (human) (Stelzl *et al.*, 2005), *Drosophila melanogaster* (fly) (Giot *et al.*, 2003)) and *Sacchromomyces cerevisiae* (yeast) (Uetz *et al.*, 2000) among other species. However, while the large number of apparent interactions appears promising at first, there is still a high false positive rate of interaction prediction in these datasets that compromises their usefulness as a standalone method for prediction (Sprinzak, Sattath & Margalit, 2003). It is suspected that the 39,000 and 40,000 protein interactions known in humans (according to the HPRD, as of September, 2012) only represent a fraction of the actual network of interactions. Experimentally testing and retesting the entire set of protein-protein pairs would be a painstakingly slow and near impossible task.

As a result, protein-protein interaction prediction has moved in a new direction over the past decade to employing computational techniques as an additional method to uncover novel potential interactions. While no method can predict interactions with 100% certainty, these prediction methods as a whole can narrow down the set of potential interactors to a subset of those most likely to be true interactions as a starting point for further lab experiments. The sections following provide an overview of the evidence considered in protein-protein interaction prediction and the methods currently available.

### 1.4.1 Evidence Incorporated into Computational Methods

The majority of computational methods employ a comparative genomics approach to predict interactions. There are four main facets of comparative genomics that have been considered in predicting protein interactions:

#### 1.4.1.1 Primary Sequence and Protein Structure

The inherent sequence and structural properties of proteins have been identified as key to mediating protein contact and potential indicators of interactions (Young, Jernigan & Covell, 1994; Sprinzak & Margalit, 2001; Bock & Gough, 2001; Chinnasamy, Mittal & Sung, 2006; Reimand *et al.*, 2012). Sprinzak et al. looked at primary amino acid sequences in yeast and found a positive correlation between pairs of signature sequences in interacting proteins that could predict interactions in two proteins with that same set of the motifs (Sprinzak & Margalit, 2001). Additionally, the physicochemical properties of amino acids, for example, charge, hydrophobicity and surface tension, have also been implemented in predicting novel interactions (Bock & Gough, 2001; Chinnasamy, Mittal & Sung, 2006). However, while each of these methods did identify probable interactions on small datasets, the specificity of what was considered only allowed limited prediction coverage, suggesting that prediction by sequence alone cannot be applied on a large scale. Additionally, attempts to predict interactions through tripeptide secondary structure motifs failed to perform significantly better than random (McDowall, 2011).

The main issue with incorporating protein structure into prediction methods is the lack of available three-dimensional structures for the majority of human proteins. However, a recently published method, PrePPI (Zhang *et al.*, 2012) (described in more detail below), has tackled this limitation by mapping the sequences of a pair of proteins to proteins with known structures that has allowed interaction models to be identified for 13,000 human proteins and has successfully predicted interactions that were verified experimentally (Zhang *et al.*, 2012).

#### 1.4.1.2 Gene Neighbouring, Co-expression and Fusion

Both the physical positioning of genes and their patterns of expression have been shown to contribute to the likelihood of their protein products interacting (Dandekar *et al.*, 1998; Teichmann & Babu, 2002; Snel, van Noort & Huynen, 2004). The link between the organisation of genes into operons in bacteria or co-regulation in eukaryotes and similarity of function has been shown in archaea (Dandekar *et al.*, 1998), prokaryotes (Dandekar *et al.*, 1998) and eukaryotes (Teichmann & Babu, 2002; Snel, van Noort & Huynen, 2004). Additionally, gene co-expression appears to be evolutionarily conserved for interacting proteins and for orthologues of interacting proteins (Hahn *et al.*, 2005; Tirosh & Barkai, 2005). However, an analysis of the relationship between gene co-expression in interacting proteins by Bhardwaj et al. found that while gene expression and interaction was correlated in *E. coli*, in yeast, mouse and human, it was not significant enough to serve as an indicator of interaction on its own (Bhardwaj & Lu, 2005). Finally, the principle of gene fusion, where two functionally-related genes are either fused into one gene or split into two genes over time, has also successfully identified protein interactions conserved across bacteria (Enright *et al.*, 1999) and novel functional associations in yeast and worm (Enright & Ouzounis, 2001).

#### 1.4.1.3 Subcellular Localisation

An intensive study comparing known protein interactions in human, yeast, fly and *Caenorhabditis elegans* (worm) by Gandhi et al. revealed that interacting proteins are more likely than not to share the same primary subcellular localisation (Gandhi *et al.*, 2006). However, this principle did not hold true in certain compartments and if one of the proteins was in a location outside of its normal environment (Gandhi *et al.*, 2006). Additionally, a recent study of protein turnover showed that despite no significant difference in protein turnover rates in the cytoplasm, nucleus and nucleolus, proteins that exist in one compartment during assembly turn over at slower rates in the compartment where the stable complex functions (Boisvert *et al.*, 2012). Overall, while shared subcellular environments can indicate potential interaction, the movement of proteins between compartments and associated change in intrinsic properties suggests that interaction prediction cannot depend on localisation alone.

#### 1.4.1.4 Orthology and Gene Co-Evolution

Orthologues, or identical genes that have remained in different species following divergence from a common ancestor, have been shown to be a useful tool in identifying and confirming protein interactions (Walhout *et al.*, 2000; Matthews, 2001; Deane, 2002). It has been suggested that for two interacting proteins in one species that share >80% sequence identity and an e-value of $<10^{-70}$ with two protein in another species, the interaction of those proteins can be inferred to occur in both species (Shoemaker & Panchenko, 2007). If the orthologues of two proteins in one species do interact in another, the pairs are referred to as ‘interologs’, or ‘interacting orthologues’.

As describe in Section 1.1 above, Gandhi et al.'s analysis of protein interactions in human, yeast, worm and fly revealed only a low number of genes (16) were conserved between the four species (Gandhi *et al.*, 2006). However, interacting orthologous pairs of proteins have been exploited on a smaller scale to map interactions between yeast and worm (Matthews, 2001) and to confirm a large set of interactions identified in yeast based on their orthologues interacting in other species (Deane, 2002). Additionally, mapping known interactions among human orthologues of yeast, worm and fly has formed the basis for construction of a putative human protein interaction map with over 71,000 interactions (Lehner & Fraser, 2004).

However, while orthologous transfer can help to identify potential interacting pairs of proteins, co-evolution of two proteins in different species does not mean that the proteins share the same function (Pazos & Valencia, 2008). Co-evolution can occur diffusely, in which many proteins with similar gene expression patterns, subcellular localisations or involved in similar biochemical pathways are affected. With widespread co-evolution, it becomes less likely that the sole reason a given pair of proteins has evolved specifically is to maintain the same function. As a result, the role of co-evolution in protein interactions is best handled by considering the specificity of pairs of co-evolved proteins within the context of the rest of the interaction network. As an alternate method, anti-correlated evolution, where one of two interacting proteins has not co-evolved, can indicate functional changes in associated pathways (Pazos & Valencia, 2008).

# 1.5 Computational Prediction Frameworks

All computational prediction frameworks consist of two components: training and testing datasets and a classification method. There are four main machine learning methods exploited by protein interaction prediction methods: Bayesian networks, support vector machines (SVMs), neural networks (NN) and random forest decision trees.

## 1.5.1 Dataset Construction

In order to train and test a prediction or classification method, two datasets, one of true positive and one of true negative entities, must be collected. Typically, these datasets are split into further subsets for cross-validation during training (see Section 1.6.2: Cross-Validation, below).

### 1.5.1.1 Positive Datasets

As shown in Table 1.2, there are multiple databases available that include varying subsets of proteins known to interact; however, as described in Section 1.3 above, the databases are all curated differently and their content does not overlap. Therefore, selection of an appropriate positive dataset resource (or multiple resources) depends upon the research being undertaken. Additionally, it cannot be assumed that any of the positive datasets are free from false positive interactions.

For humans, the HPRD provides the largest single source of protein-protein interactions with 41,327 interactions manually curated from literature included (as of October 2012) (Prasad, n.d.). In addition to interaction information, the HPRD also includes features

for identifying phosphorylation motifs (PhosphoMotif Finder), information about human signalling pathways via NetPath and annotations on protein isoforms, post-translational modifications, subcellular localisations and enzyme-substrate relationships. Finally, the HPRD has also recently introduced the 'Human Proteinpedia' to allow researchers to input directly experimental data about identified interactions that can link to their corresponding entries in the database (Prasad, n.d.).

#### 1.5.1.2 Negative Datasets

Assembling a true negative dataset is more complicated as there is no extensive, centralised source of negative interactions. Currently, the only negative interaction resource that exists is the Negatome (Smialowski *et al.*, 2010) (for the 'negative interactome'), a small database with approximately 1000 pairs of proteins shown not to interact that have been curated from literature. However, the Negatome represents only a fraction of the number of true negative interactions that is not large or diverse enough to be used as a complete negative dataset. Instead, negative interaction datasets are constructed by two main methods.

In the first method of negative dataset construction, pairs of proteins are selected as negative interactions on the basis that they exist in different subcellular locations within the cell (e.g. one protein is localised to the plasma membrane and the other is located in the nucleus) (Jansen *et al.*, 2003; Rhodes *et al.*, 2005; Xia, Zhao & Huang, 2010). However, as addressed above (in Section 1.4.1.3: Subcellular Localisation), proteins can be present in different compartments depending on their state and function at a given time (Boisvert *et al.*, 2012). Additionally, selecting negative datasets in this manner can

positively bias a prediction method if it incorporates subcellular localisation into its analysis (Ben-Hur & Noble, 2006).

As a result, a second method of dataset construction selects protein pairs at random and then filters out any interactions that have been previously annotated as positive in all or individual interaction databases (Scott & Barton, 2007; Qi, Klein-Seetharaman & Bar-Joseph, 2007). With an estimated ratio of one in thousands of potential protein pairs in humans thought to interact, randomly selecting pairs still ensures that over 99.8% are true non-interactors (Qi, Klein-Seetharaman & Bar-Joseph, 2007).

### 1.5.2 Cross-Validation

One of the barriers of effective machine learning methods is over-fitting the method such that it is capable of recognising the examples it was trained on but not the unseen examples in a test set. The cross-validation method of learning attempts to minimise this potential bias. Figure 1.4, below, provides a schematic description of *k*-fold cross-validation. In this method, a dataset is divided into *k* number of subsets. During each of *k* rounds, the predictor is trained on all but one of the subsets and tested on the remaining set. With each round, the training and testing sets rotate such that each round is tested on a different subset until all subsets have been the test set. As a variation, the 'leave one out' method, extends *k*-fold cross-validation such that *k* is the total number of samples and training is repeated for *k* rounds on all samples but one until every sample has been the test.

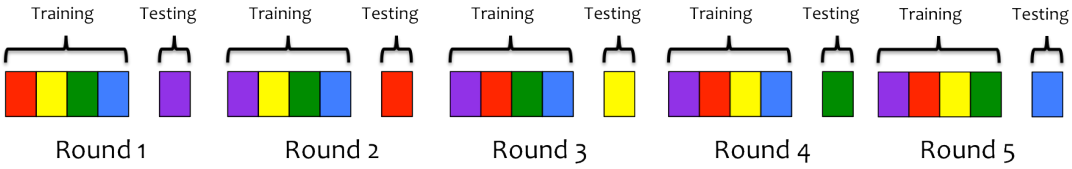


**Figure 1.4: Schematic of the Cross-Validation Training and Testing Process.** *K*-fold (in this case, *k*=5) cross-validation dataset separation for training and testing the computational prediction methods is depicted schematically. During each round, the predictor is trained on four of the training subsets (marked by the braces as 'training') and tested on one subset (marked by the braces as 'testing'); the sets then rotate such that each subset is made the test set once.

### 1.5.3 Measuring Prediction Accuracy and ROC Plots

While there are several ways of measuring prediction accuracy, one method involves comparing how well a predictor can identify true positive interactions and if it over-predicts negative interactions as positive (Metz, 1978). These two values, the True Positive Rate (TPR or sensitivity, Equation 1.1) and False Positive Rate (FPR or specificity, Equation 1.2), respectively, are calculated at a range of prediction cut-off thresholds to give values between 0.0 and 1.0 that are plotted against each other as a Receiver Operator Characteristic (ROC) curve.

$$TPR = \frac{TP}{TP+FN}$$

**Equation 1.1: True Positive Rate.** The True Positive Rate (TPR), or Sensitivity, is calculated by dividing the number of true positive predictions by the total number of positives, where TP is true positives and FN is false negatives.

$$FPR = \frac{FP}{FP+TN}$$

**Equation 1.2: False Positive Rate.** The False Positive Rate (FPR), or Specificity, is calculated by dividing the number of false positive predictions by the total number of negatives, where FP is false positives and TN is true negatives.

When plotted with the FPR on the x-axis and the TPR on the y-axis, the ideal ROC plot would form a smooth curve starting from the lower left corner of the graph (0, 0) and climb to the upper right corner of the graph (1, 1) with a maximum amount of space enclosed by the curve and the x- and y-axes (shown in Figure 1.5 and labelled A). Likewise, a ROC curve that falls on or below a straight line bisecting the plot from (0, 0) to (1, 1) indicates that predictions are worse than random (shown as B and C). While comparing the curve trajectories of ROC plots of different methods can show which perform better than others, a quantitative measure of the accuracy can be obtained by calculating the respective areas under the curves (AUC), a value representing the percentage that if an example is randomly selected, it will have been assigned the correct prediction or output value.

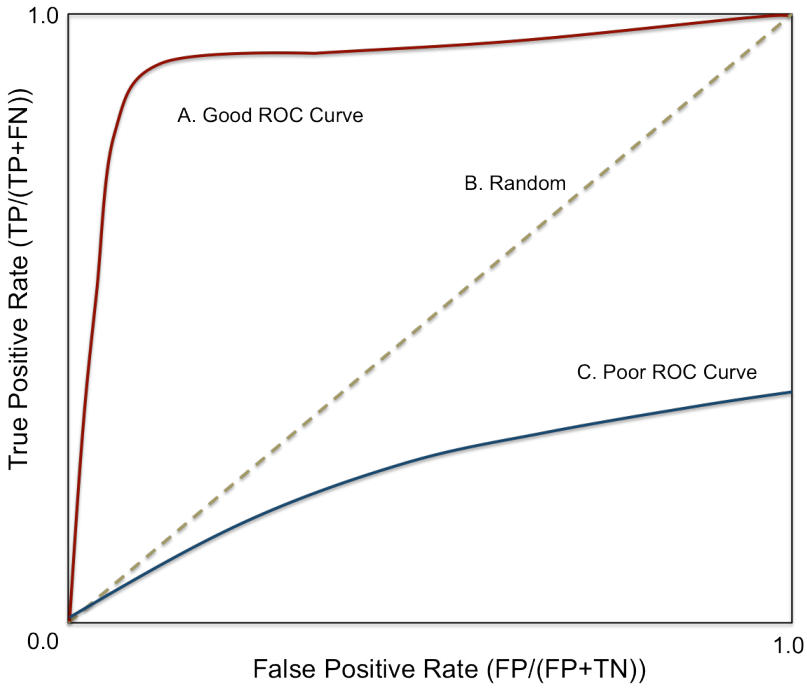


**Figure 1.5: Examples of good and poor ROC curves.** A) A good ROC curve starts at (0, 0) and maximises the area beneath the curve. B) The line bisecting the plot from (0, 0) to (1, 1) indicates the curve where the true and false positive rates are equal at all threshold and prediction are random. C) A poor ROC curve falls anywhere below the random line and has a low area below.

Another accuracy measure, Matthew's Correlation Coefficient (MCC) or the *Phi* Coefficient, calculates the correlation between expected and observed predictions (Equation 1.3). Low MCC values (i.e. between -1.0 and 0.0) indicate no correlation between expected and predicted outcomes, while high MCC values (i.e. between 0.0 and 1.0) indicate a full, positive correlation between what was expected and what was predicted. Comparison of the MCC statistics for different predictors can therefore provide an additional measure of performance.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

**Equation 1.3: Matthew's Correlation Coefficient.** Matthews's Correlation Coefficient (MCC) measures the correlation between expected and observed predictions, where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives.

Other forms of the ROC plot, the ROC50 or ROC100 curves, plot the number of false positives on the x-axis versus the number of true positives on the y-axis to compare the absolute numbers of the highest scoring positive predictions that are correctly predicted before the highest scoring 50 (or 100) negatives are incorrectly predicted as positive. These true positive versus false positive plots offer a more detailed measure of how well different methods perform for the highest scoring predictions. An ideal plot would show a curve with a vertical increase from (0, 0), indicating that a high number of true positives were predicted before the first false positive results were returned.

## 1.5.4 Learning Methods

### 1.5.4.1 Bayesian Methods

The Bayesian Network learning method, first described in 1764, involves calculating the probability that an event would occur given evidence from prior events. The method centres around Bayes' Theorem (Equation 1.4):

$$P(h|E) = \frac{P(h)P(E|h)}{\sum_{i=1}^{n} P(E|h_i)P(h_i)}$$

**Equation 1.4: Bayes' Theorem.** Bayes' Theorem calculates the final probability that a hypothesis ($h$) is true based on a given set of evidence ($E$) as the probability of $h$ multiplied by the probability of $E$ given $h$ (P($h$)P($E$|$h$)) divided by the sum of the probabilities $E$ being true given $h$ multiplied by the probability of $h$ for each case test case.

In a two-case classification scenario where the result can either be 'true' or 'false', Bayes' Theorem calculates the posterior odds ratio (P($h$|$E$)), or probability that the hypothesis will be true given a new piece of evidence. This calculation requires knowledge of the prior odds ratio, or the estimate of the likelihood that if an example from the entire set of examples is chosen at random before assessing the hypothesis (P($h$)), it will be true. For example, if 5 out of 25 total examples were true, the prior odds ratio would be equal to 5/25 or 1/5, meaning that if a piece of evidence were chosen at random, there would be a 1/5 chance that it would be true.

The product of the probability that the test example will be true and the prior odds ratio are then divided by the sum of the probabilities that the example could be each of the hypothesis results. When expanded for a true/false classification, the denominator of the theorem's equation is calculated by multiplying the probability that the piece of

evidence may be true by the prior odds ratio and adding it to the probability that the piece of evidence may be false multiplied by the prior odds ratio of selecting a false example (i.e. 1-prior odds ratio) (Equation 1.5).

$$P(h|E) = \frac{P(E|h) \times P(h)}{P(E|h) \times P(h) \times P(E|\neg h) \times P(\neg h)}$$

**Equation 1.5: Expanded Bayes' Theorem for a True/False Classification Example.** P($E|\neg h$) and P($\neg h$) represent the probability that the piece of evidence is false given the hypothesis is false and the prior odds ratio that the hypothesis is false.

Overall, full Bayesian classifiers work under the assumption that individual pieces of evidence are co-dependent and that the probability of the theory being true is influenced by the combined probabilities of these evidences.

#### 1.5.4.2 Naïve Bayesian Classifiers

Bayesian networks can be further simplified into naïve Bayesian classifiers that are based on the assumption that each individual piece of evidence is independent. With this assumption, Bayes' Theorem can be rewritten (Equation 1.6).

$$P(E_1, E_2 ..., E_n|h) = P(E_1|h)...P(E_n|h)$$

**Equation 1.6: Naïve Bayes' Theorem.** Evidence set *E={E1, E2, ..., En}* includes *n* evidence.

The posterior probability for the entire evidence is computed by multiplying the individual probabilities of each piece of evidence. Because naïve Bayesian classifiers allow individual sources of evidence to be considered one at a time, calculation of the ultimate probability is quicker.

### 1.5.4.3 Artificial Neural Networks

#### 1.5.4.3.1 General Overview

Artificial Neural Networks (ANNs or 'neural networks') are computational architectures that process and learn from information in a method analogous to the biological networks in the human brain (Hecht-Nielsen, 1987; Azam, 2000).

Biological neural networks are composed of an intricate network of billions of interconnected neurons (Gurney, 1997). At its most basic, a biological neuron is made up of a cell body, or 'soma', with small protrusions, called 'dentrites', and one large extension, called the 'axon', branching off of it and connecting with other adjacent neurons. When the neuron receives a stimulatory signal or perceives information from its surrounding environment, it sends an impulse, or 'action potential', down its axon shaft that in turn either positively excites or negatively inhibits the neurons adjacent to and in contact with it. Over time, repeated presentation of stimulatory information to the neurons causes them to learn how strong, where and when these action potentials should fire, weighting the connections such that there is a specific response to a stimulus (Gurney, 1997).

While the above description is an over-simplification of the fine details of biological neural networks, artificial neural networks are built upon the same principle of interconnected neurons (or 'nodes') that process input information to determine an output based upon what they have learned from previously processed inputs. The most basic ANN, the perceptron, consists of two layers: an input layer that receives the stimulus information for distribution and an output layer that processes the information

from the input layer and determines an appropriate outcome (ROSENBLATT, 1958). However, most ANNs also contain one or more additional, hidden layers to help with the information processing and classification and are therefore multi-layer perceptrons (MLPs).

How neural networks process input data is dependent upon the combination of the network structure and how it has learned to handle patterns. The neurons in the three network layers can be either partially or wholly interconnected; each of these connections is assigned a weight, a value between -1 and +1, that corresponds to how strong the passage of information to and from the two connected nodes should be. Initially, the network is constructed with the nodes linked by randomly assigned weight values. Then, through the successive presentation of a series of training patterns, the network 'learns' how to handle the data to achieve a desired outcome (in supervised networks) or an outcome most consistent with the other patterns (in unsupervised networks) and repeatedly readjusts the connection weights to minimise the error between the calculated and expected outcomes. The exact method of this error incorporation and weight adjustment depends upon the learning function of the specific networks. However, all neural networks are trained with the ultimate goal of reaching a point where the error is as low as possible, stable and presentation of new or imprecise patterns can still produce the expected output.

Fundamentally, ANNs are based upon three main assumptions (Azam, 2000):

*Generalisation* - The network will provide general outcomes based upon what it has learned from previous cases.

*Degradation* - Performance of the neural network will decrease if the data presented to it is incomplete or imprecise.

*Adaptation/Learning* - The neural network will gain knowledge from the data it is presented with and will try to maintain that knowledge throughout all of training.

#### 1.5.4.3.2 Feed Forward Neural Networks

Feed forward learning methods process information in one direction (from the input to hidden to output layers) and is implemented in most of the well-known algorithms, including back propagation, self-organising maps, Kohonen networks and adaptive resonance networks, among others. Though there are other methods of processing that involve carrying information both backward and forward, these methods have not been applied in this thesis and are not discussed. In applying neural networks to protein-protein interaction prediction, we have focused on two main methods: Back propagation and Scaled Conjugate Gradient (SCG).

#### 1.5.4.3.3 Back Propagation

In the Back propagation network algorithm, the most widely employed learning function due to its flexibility for use in both simple and complex network structures, training occurs in two successive steps: a 'forward pass' and a 'backward pass' (Hecht-Nielsen, 1990).

In the first, 'forward pass' step, input is transferred through the layers until the output layer is reached and a final output activation is calculated. In a three-layer, multi-layer perceptron, the input layer consists of one or more neurons, or 'nodes', that receive

information as a 'pattern' with one value per input node. Each node within the input layer is linked to one or more nodes in the hidden layer by weighted connections. During processing, the node that the output from input node should be directed to is determined through a two-step process. First, the activation ($a_i$) for the input node is computed by taking the sum of the weighted connections between all nodes connected to it and factoring in any bias (Equation 1.7):

$$a_i = \sum_j w_{ij} o_j + b_i$$

**Equation 1.7: The activation of an input node.** The activation for each unit where $b_i$ is the bias (the connection weight from a node consistently having an output of 1.0), $w_{ij}$ is the connection weight between the nodes $i$ and $j$ and $o_j$ is the output for node $j$.

Then, taking into account this activation value, the output for the input node is then calculated (Equation 1.8):

$$o_i = \frac{1}{1 + e^{-a_i}}$$

**Equation 1.8: The output.** The output for input node i ($o_i$) where $a_i$ is the activation of that node (see Equation 1.7).

These calculation are then repeated between the hidden and output layers to give a final output value.

After one 'cycle' (or 'epoch') in which the network has seen the entire set of patterns, the second, 'backward pass' step occurs. In this step, the total error for the cycle is

calculated by summing the difference between the resulting output and the expected output for each input pattern (Equation 1.9):

$$E = \sum_p E_p = \sum_p (x_p - o_p)^2$$

**Equation 1.9: Total Error.** Calculation of the total error ($E$) for one training cycle as the sum of the individual error for each training pattern, where $E_p$ is the error for one, individual input pattern, $x_p$ is the expected output for the pattern and $o_p$ is the calculated output for the pattern.

After determining this total error, the weights for each connection are re-adjusted in the reverse order from output to hidden to input nodes. This change in weight is calculated by dividing the derivative of the total error (E) by the derivative of the current weight of the connection ($w_{ij}$) and multiplying the result by the negative of the learning rate ($k$), a constant ranging from 0.0 to 1.0 that determines how much the connection is re-weighted (Equation 1.9).

$$\Delta w_{ij} = -k \frac{\delta E}{\delta w_{ij}}$$

**Equation 1.9: Change in connection weight.** Calculation of the change in weight for the connection between nodes $i$ and $j$, where $k$ is the learning rate and $E$ total error for one cycle.

This cycle of forward training and backward correction repeats, with each pattern presented until the error has converged at a set minimum value or a determined number of cycles of training has been completed, allowing the network to 'learn' which outputs respond to different inputs (Hecht-Nielsen, 1990; Basheer & Hajmeer, 2000).

#### 1.6.4.3.4 Scaled Conjugate Gradient

The Scaled Conjugate Gradient (SCG) method processes information through a first 'forward pass' step similar to the back propagation methods; however, it diverges in the second, 'backward pass' step in how it updates the connection weights between nodes.

While in back propagation, connections are re-weighted down the error gradient such that each error minimisation is based only upon the specific cycle of training, conjugate methods update connection weights to build upon the previous iteration by accounting for the existing weights and error calculations (Moller, 1990). By not partially undoing the learning completed in previous steps, conjugate gradient methods are able to arrive more quickly and accurately at a local error minimum. Additionally, while back propagation methods are highly reliant on user-supplied parameters, namely for the learning rate (how much weights should be updated) and the momentum coefficient (how often weights are updated), these values are encompassed in the SCG algorithm itself, allowing for less possibility for user-error.

#### 1.6.4.3.5 Network Architecture

The structure of a neural network depends upon the learning function implemented, the data presented and the output desired. The number of input nodes is determined by the composition of the set of data in the input pattern (called a 'pattern' or 'vector') and is typically arranged such that there is one node per input value, with each value normalised to between 0.0 and 1.0 in order to eliminate any bias caused by inconsistent ranges. The number of output nodes corresponds to the goal of what the network is being employed for. For example, a simple neural network looking to predict whether a

case is one of two possibilities might only have one output node (a value standing for either ‘true’ or ‘false’), while a more complex network determining which letter a particular image shape is most like might have 26 different output nodes (with one for each letter).

The number of hidden nodes within the network varies widely, and while a number of different methods for determining the optimal number of nodes have been proposed, the number for specific network is typically best determined through trial and error through a wide range of options (Basheer & Hajmeer, 2000). The number of hidden nodes should not be too low such that all patterns are immediately grouped together but also not too high that all noise in the data is treated as a distinct case so that the network cannot classify input patterns properly (Basheer & Hajmeer, 2000).

Additional parameters exist for the specific learning functions that can determine, for example, the rate at which the network learns, how fast weights are updated and when weights are updated, and are often set through trial and error in order to optimise the network for minimal error.

#### 1.5.4.5 Additional Learning Methods

Two other main machine learning methods are employed in protein interaction prediction: Support Vector Machines (SVMs) (Bock & Gough, 2001; Martin, Roe & Faulon, 2005; Hue *et al.*, 2010; Xia, Zhao & Huang, 2010; Zaki *et al.*, 2011) and Random Forest Decision Trees (Breiman, 2001; Qi, Klein-Seetharaman & Bar-Joseph, 2005; Qi, Bar-Joseph & Klein-Seetharaman, 2006). Briefly, SVMs classify data into one of two categories by rearranging data into two cases based on mathematical

functions (called 'kernels'). This separation creates a division (called a 'hyperplane') where, in the case of protein-protein interaction, pairs that fall on one side are considered 'interacting' and on the other 'non-interacting'. In Random Forests, each piece of evidence is analysed by multiple decision trees. When the sample is assigned a classification by each tree in the 'forest', that classification receives a 'vote'. After the sample has been analysed by all of the trees, the votes are tallied, and it receives the outcome for the classification with the most votes. SVMs and Random Forests are out of the scope of this thesis and will not be discussed in detail.

### 1.5.5 Current Human Protein-Protein Interaction Predictors

#### 1.5.5.1 STRING (http://string-db.org)

The STRING interaction database incorporates both known and predicted interactions for 1100 archaeabacteria, prokaryotic and eukaryotic organisms (Mering *et al.*, 2005; Szklarczyk *et al.*, 2011). Rather than predicting only direct, physical interactions, the STRING prediction method focuses on identifying functional associations between proteins and creating a network of the entire set of possible interactions. Predictions are based upon the combination of relevant genomic context information (i.e. conserved gene neighbourhoods, gene fusion events and co-occurrence of genes across genomes) and the sharing of gene functions across proteins belonging to the same orthologous groups. Additionally, STRING assigns a confidence score to both predicted and known interactions based on if the two proteins are found in the same biological pathway according to the KEGG database. The STRING web interface allows uses to visualise

both predicted and known interactions in an interactive network viewer that allows detailed viewing of the data supporting connections (Szklarczyk *et al.*, 2011).

#### 1.5.5.2 OPHID/I2D (http://ophid.utoronto.ca/ophidv2.201/)

I2D, formerly OPHID, contains predicted interactions derived from mapping high-throughput data for interacting proteins in one species to another (Brown & Jurisica, 2005). The original OPHID prediction set was generated by the reciprocal best-hit approach by matching proteins in known protein-protein interactions to their human orthologues through BLAST. In recent years, OPHID has been extended to include predicted interactions in yeast, worm, fly, mouse and Rattus norvegius (rat) by transferring human interactions to their orthologues in each species (Brown & Jurisica, 2007). The I2D website provides both known and predicted interactions.

#### 1.5.5.3 FunCoup (http://funcoup.sbc.su.se/)

FunCoup (for 'functional coupling') is a modified naïve Bayesian protein-protein interaction predictor that incorporates evidence from phylogenetic profiles, subcellular localisation, known protein-protein interactions, mRNA co-expression, shared transcription factors, co-microRNA regulation and domain associations to predict how likely two proteins are to be functionally coupled (Alexeyenko & Sonnhammer, 2009). Protein pairs from human, yeast, worm, fly, mouse, rat and *Arabidopsis thaliana* (plant) are split into four classes depending on their functionality (signalling pathways, metabolic pathways, known protein-protein interactions and members of the same complex) and then trained on each to give a raw score that is allocated to a bin and assigned the score associated with that bin. While FunCoup primarily deals with

matching proteins by function rather than direct interaction, the resulting network produced is similar to those resulting from direct interaction prediction methods.

#### 1.5.5.4 IntNetDB v. 1.0 (http://hanlab.genetics.ac.cn/sys/)

IntNetDB v. 1.0 is a Bayesian interaction predictor for human protein-protein interactions (Xia, Dong & Han, 2006). Predictions are computed by calculating the probability that two proteins will interact based on seven pieces of evidence: gene co-expression, genetic interactions, phenotype similarity, GO term similarity, domain-domain interactions and gene context.

## 1.6 PIPs: A Predictor of Human Protein-Protein Interactions

Over the past six years, our group has developed a novel protein-protein interaction predictor (PIPs) (Scott & Barton, 2007; McDowall, Scott & Barton, 2009). PIPs utilises a naïve Bayesian framework to determine the likelihood that a pair of proteins will interact based on evidence from a range of features (Scott & Barton, 2007; McDowall, Scott & Barton, 2009; McDowall, 2011). Unlike in neural network and SVM machine learning methods, where a vector of scores is fed into a computational 'black box' to produce a final outcome, the output generated by PIPs includes a breakdown of the individual scores for each source of evidence. This breakdown is useful for further investigation into both positive and negative predictions.

### 1.6.1 The PIPs Framework

Since its inception, PIPs has undergone two prior version releases. Table 1.3 and Figure 1.6, below, compare the differences between the algorithms of versions 1.0 (A) and 2.0 (B). In both versions, the first stage of the method considers six individual pieces of evidence split in three 'modules' ('Expression', shown in red, 'Orthology', shown in yellow, and 'Combined', shown in green, described in more detail in section 1.6.2). While the Expression and Orthology modules, which consider mRNA co-expression patterns and orthologous interactions, respectively, are consistent in both versions, the Combined module was re-engineered in v. 2.0 to include GO term similarity in place of subcellular localisation (McDowall, 2011).

In the first stage of PIPs, each of the Expression, Orthology and Combined modules is trained independently on a set of positive and negative interactions. During training of each module, a pair of proteins is assigned a score based on its available evidence. This score places it in a 'bin' covering a range of scores appropriate to the module. For example, the Expression module contains 20 bins, with each covering a 0.1 range such that every score from -1.0 to 1.0 is covered by a bin. If a protein pair has an expression score of 0.93, it would be assigned the bin covering scores from 0.8 to 0.9. Once all of the pairs in the training set have been assigned a bin, the numbers of positive and negative pairs are counted. To calculate a likelihood ratio for each bin, the number of positives assigned to the bin is first divided by the total number of positive pairs being considered. This value is then divided by the number of negatives assigned to the bin divided by the total number of negative pairs being considered (Equation 1.10).

Following training, each bin for the module has been assigned a specific likelihood ratio:

$$LR = \frac{Pos_{Bin} / Pos_{Total}}{Neg_{Bin} / Neg_{Total}}$$

**Equation 1.10: Likelihood ratio calculation.** The likelihood ratio for each bin is calculated by dividing the number of positive pairs assigned to the bin by the total number of positive pairs that could have been assigned to the bin and then dividing the value by the number of negative pairs assigned to the bin by the total number of negative pairs that could have been assigned to the bin.

During testing and prediction, a protein pair is allocated to a bin based on its evidence score for each module and then assigned the likelihood ratio associated with that bin. A combined likelihood ratio ($LR_{EOC}$) for the first stage is then calculated for each pair by multiplying the likelihood ratios for the Expression, Orthology and Combined modules together. Pairs with $LR_{EOC}$ values above a set threshold are then assembled into a preliminary interaction network.

The second stage of PIPs considers as evidence this preliminary interaction network in either the Transitive module (T) (introduced in v. 1.0 and included in v. 2.0, shown in purple) or the Cluster module (M) (introduced in v. 2.0, shown in blue). Like the Expression, Orthology and Combined modules, likelihood ratios are assigned to each bin during training, and protein pairs are assigned the likelihood ratio associated with the bin it is allocated to during testing and prediction. As one of the assumptions of the naïve Bayesian network is the independence of each source of evidence (or 'module', as referred to in PIPs) considered, the Transitive and Cluster modules, which both take the

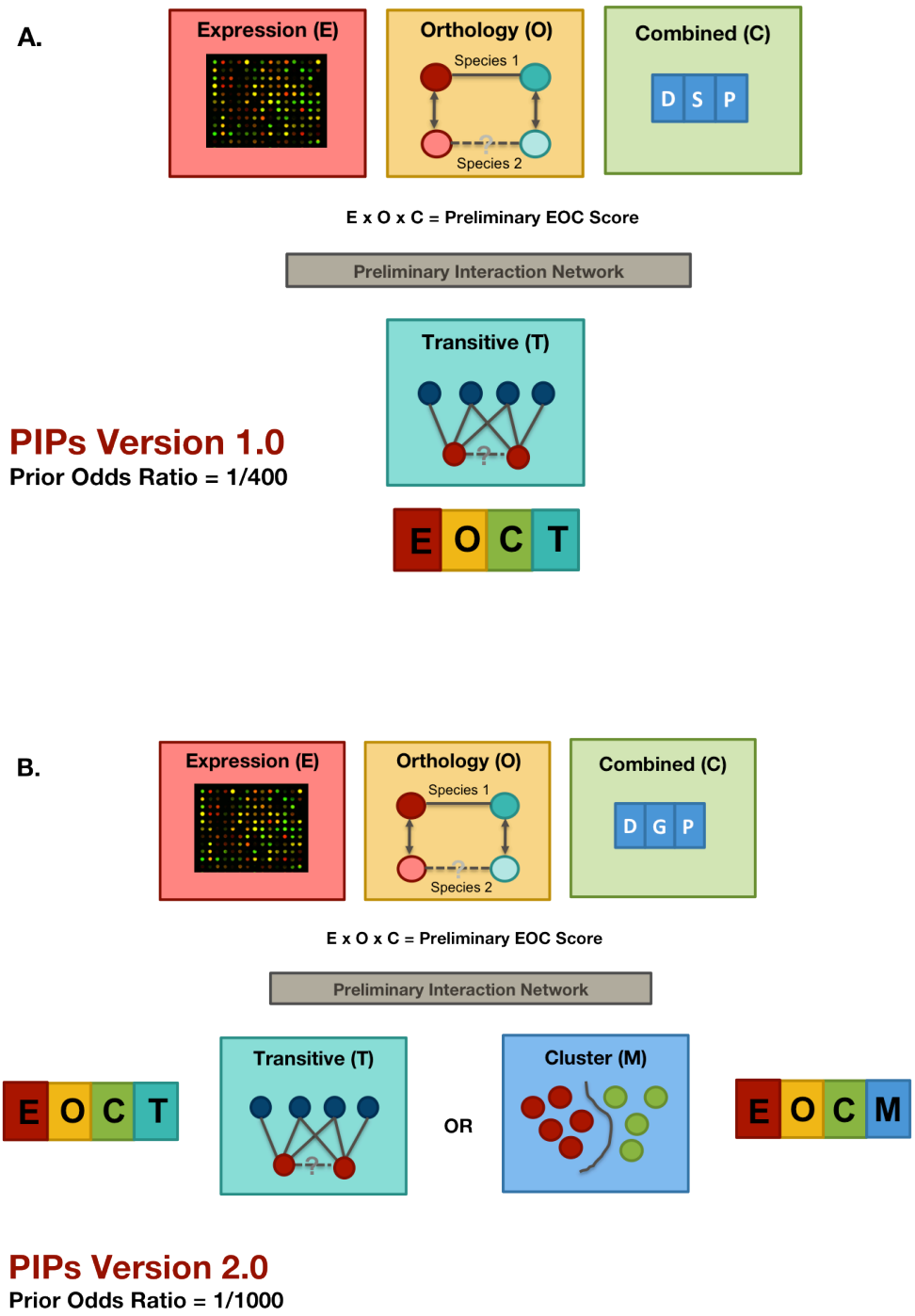


**Figure 1.6: Schematic diagram of PIPs v. 1.0 and v. 2.0.** The algorithm and details of the modules included in PIPs v. 1.0 (A) and PIPs v. 2.0 (B) are shown. Three main developments were made to the predictor for v. 2.0: 1) in the Combined module, the Subcellular Localisation component was removed and replaced with GO Term Similarity; 2) the Cluster module was added as a second option for network analysis in Stage II; and 3) the prior odds ratio was changed from 1/400 to 1/1000 to reflect more accurately the number of positive interactions, altering the cut-off threshold for prediction from 400.0 to 1000.0.

| | A. Version 1.0 | B. Version 2.0 |
|---|---|---|
| **Development Details** | Michelle Scott (Scott & Barton, 2007) | Mark McDowall (McDowall, 2011) |
| **Expression** | **Data Source**: GDS596 from the Gene Expression Omnibus<br>**Scoring Method**: Pearson's Correlation Coefficient<br>**Number of Bins**: 20 | **Data Source**: Microarray E-GEOD-7307 (A-AFFY-44 microchip)<br>**Scoring Method**: Pearson's Correlation Coefficient<br>**Number of Bins**: 20 |
| **Orthology** | **Data Source**: InParanoid, BIND, DIP and GRID databases<br>**Scoring Method**: Organism-based using InParanoid score and known yeast, worm and fly interactions<br>**Number of Bins**: 13 | As in v. 1.0. |
| **Combined** | **Components**: Domain Co-Occurrence, PTM Co-Occurrence, Subcellular Localisation<br>**Data Source**: Domains (InterPro and PFAM), PTMs (HPRD and UniProt), Subcellular Localisation (PSLT, human subcellular localisation predictor)<br>**Scoring Method**: Domains (Chi-squared score), PTMs (PTM Co-occurrence score), Subcellular localisation (qualitative score for proximity of components)<br>**Number of Bins**: Domains (5), PTMs (4), Subcellular Localisation (4) | **Components**: Domain Co-Occurrence, PTM Co-Occurrence, GO Term Similarity<br>**Data Source**: Domains (InterPro and PFAM), PTMs (HPRD and UniProt), GO Term Similarity (Gene Ontology Database)<br>**Scoring Method**: Domains (Chi-squared score), PTMs (PTM Co-occurrence score), GO Term Similarity (Jiang and Conrath's method with GraSM adjustment)<br>**Number of Bins**: Domains (5), PTMs (4), GO Term Similarity (3) |
| **Transitive** | **Data Source**: Preliminary interaction network predicted from Expression, Orthology and Combined modules (cut-off $LR_{EOC}$ = 10.0)<br>**Scoring Method**: Neighbourhood topology; Counts common edges/interactions shared between proteins in a pair.<br>**Number of Bins**: 5 | As in v. 1.0. |
| **Cluster** | Not in v. 1.0. | **Data Source**: Preliminary interaction network predicted from Expression, Orthology and Combined modules (cut-off $LR_{EOC}$ = 5.0)<br>**Scoring Method**: Markov Clustering Algorithm (MCL); Proteins 'clustered' into groups by network analysis.<br>**Number of Bins**: 6 |

**Table 1.3: Comparison of PIPs v. 1.0 and v. 2.0.** Details of the differences between PIPs v. 1.0 and v. 2.0 are provided. For each module, the data source, scoring method and number of bins is given. Further details of the modules are provided in Section 1.6.2, below.

same preliminary interaction network as input, cannot be included in the same final PIPs predictor. As a result, PIPs v. 2.0 includes two prediction methods: the EOCT predictor, which calculates its final likelihood ratio by multiplying the initial $LR_{EOC}$ value by the likelihood ratio assigned to the pair for the Transitive module, and the EOCM predictor, which calculates its final likelihood ratio by multiplying the $LR_{EOC}$ value by the likelihood ratio assigned to the pair for the Cluster module.

As the final step to prediction, a posterior odds ratio is calculated for each pair of proteins by multiplying the final EOCT or EOCM score by the prior odds ratio, or the probability that a protein pair will interact if it were selected at random from the set of potential positive pairs. Since there is no complete set of known interacting and non-interacting proteins, the prior odds ratio is an estimate. The prior odds ratio was originally set in v. 1.0 at 1/400 based on estimates for yeast and was revised to 1/1000 in v. 2.0 to more accurately reflect the number of human protein-protein interactions (Scott & Barton, 2007; McDowall, 2011).

## 1.6.2 Details of the PIPs Modules

### 1.6.2.1 Expression (E)

The Expression (E) module is based on the principle that if the genes are co-expressed, their protein products are more likely than not to interact (addressed in more detail in Section 1.4.1.2: Gene Neighbouring, Co-Expression and Gene Fusion, above). As evidence, the Expression module considers the Pearson's Correlation Coefficient calculated for the pair based on an mRNA microarray dataset.

#### 1.6.2.2 Orthology (O)

The Orthology (O) module is based on the principle that if two proteins interact in one species, their orthologous human proteins, if applicable, are more likely than not to interact (described in more detail in 1.4.1.4: Orthology, above).

The module relies on evidence provided by the InParanoid database, which compares the sequences between proteins in different species to identify and score orthologues (Remm, Storm & Sonnhammer, 2001; O'Brien, Remm & Sonnhammer, 2005; Ostlund *et al.*, 2010). To designate orthology groups, InParanoid starts with two seed orthologues that are identified by NCBI BLAST-searching the proteomes of two different species for pairwise best hits. These groups are then expanded by adding 'inparalogs', or proteins that are more similar to the seeds than to other sequences within the proteome, giving two main groups of inparalogs within the orthology group (one for the seed of the first species and one for the seed of second species). The inparalogs are then clustered into non-overlapping groups. An InParanoid score is then calculated for each member of each cluster that ranges from 0.0 to 1.0, corresponding to how far the cluster member is relatively from that cluster's inparalog-seed pair compared to how far the two seed orthologues are from each other. Scores of 1.0 represent a distance identical to the distance between the original seed orthologues, while scores of 0.0 represent a distance identical to the distance between the inparalog-seed pair of the cluster (Remm, Storm & Sonnhammer, 2001).

Using these orthologues, PIPs compares interactions between orthologous proteins in yeast, worm, fly and human, such that protein pairs with a known interactions in other species are assigned bins based on their InParanoid score and the number of interologs

that are found. For protein pairs with recognised interologs, the Orthology module has proven to be a strong predictor of positive interactions (McDowall, 2011).

#### 1.6.2.3 Combined (C)

The Combined (C) module considers three sources of evidence: post-translational modification co-occurrence, domain co-occurrence and GO term similarity. For post-translational modifications (PTM) component, a score is assigned to each protein pair based on how many, if any, of the pairs of PTMs for the pair are also seen in protein pairs known to interact (Equation 1.11).

$$PTM_{score} = \frac{P\left(PTM[i], PTM[j] | I\right)}{P\left(PTM[i] | I\right) \times P\left(PTM[j] | I\right)}$$

**Equation 1.11: PTM co-occurrence score.** A score for co-occurrence is calculated for a pair of post-translational modifications (*PTM[i]* and *PTM[j]*) by dividing the probability that the pair of PTMs is seen within the set of all interacting proteins (*I*) by the probability that each is seen separately.

In the domain co-occurrence component, the InterPro and PFAM domains for each pair of protein are assigned are assigned a Chi-squared score based on how often they are seen in proteins known to interact.

In v. 1.0, the third component considered subcellular localisation (Scott & Barton, 2007); however, in v. 2.0, the feature was removed and replaced with Gene Ontology (GO) term similarity (McDowall, 2011). Gene Ontology (GO), a hierarchical vocabulary of terms assigned to genes and gene products, contains three branches describing Molecular Functions (F), Cellular Compartment (C) or Biological Process (B) (Ashburner & Lewis, 2002; Harris *et al.*, 2004) For each branch, the terms are

organised by semantic similarity into a Directed Acyclic Graph (DAG), and it is hypothesised that the GO terms for interacting proteins are located closely on the DAG. Although there are many options for measuring semantic similarity, the GraSM adjustment (Couto, Silva & Coutinho, 2007) combined with the Jiang and Conrath's measure for semantic similarity (Jiang & Conrath, 1997) was chosen.

Assigning semantic similarity between two proteins involves a multi-step calculation. First, the frequency that a parent term and all of its child terms appear is calculated by counting the number of times that a term appears within entire GO hierarchy (Equation 1.12).

$$Freq(t) = Count(t) + \sum_{i \in C_t} Count(t_i)$$

**Equation 1.12: Frequency of a parent term and its child terms in the GO hierarchy.** The frequency of a term $t$ is the number of times the term (*Count(t)*) and the set of its children terms $C_t$ appear in the GO hierarchy.

Next, the probability (*P(t)*) that a term will occur within a specific branch of the GO hierarchy is calculated (Equation 1.13).

$$P(t) = \frac{Freq(t)}{\max Freq}$$

**Equation 1.13: Probability of a term being in a specific branch of the GO DAG.** *Freq(t)* is the frequency that the term appears within the branch and *maxFreq* is the frequency that term *t* appears in all branches of the hierarchy.

Taking the negative log of the probability gives the information content (*IC*) of a term (Equation 1.14)

$$IC(t) = -\log(P(t))$$

**Equation 1.14: Information Content of a term.** *IC(t)* is the information content of a specific term and *P(t)* is the probability that the term will appear in a branch in the GO hierarchy.

Finally, the semantic similarity is calculated as the information content of the most common ancestor between two terms. If two terms have multiple common ancestors, *Share*, the *IC* value of a common ancestor between terms $t_1$ and $t_2$, is calculated as the average of all common disjunctive ancestors for the term (Equation 1.15).

$$Share_{GraSM}(t_1, t_2) = \overline{IC(a) | a \in CommonDisjAnc(t_1, t_2)}$$

**Equation 1.15: Calculation of *Share*.** *Share* is the information content value of a common ancestor between two terms ($t_1$ and $t_2$) where *a* is the average of all common disjunctive ancestors for the term.

The final semantic distance between two terms is the difference between the sum of the information contents of those two terms and two times the information content of the most common ancestor (Jiang & Conrath 1997; Couto et al. 2007).

$$Share_{ICGraSM}(t_1, t_2) = \frac{1}{IC(t_1) + IC(t_2) - 2 \times Share_{GraSM}(t_1, t_2)}$$

**Equation 1.16: Final semantic similarity.** The final semantic similarity ($Share_{GraSM}(t_1, t_2)$) is the sum of the information contents (*IC*) for each term minus two times the *Share* value (Equation 1.15) for each term.

After testing which of a range of potential combinations of one or more of the three branches in the GO hierarchy was most effective in aiding correct interaction prediction,

the Biological Process branch was selected for inclusion in the Combined module (McDowall, 2011).

After each of the PTM, domain and GO term components are considered, the pair is assigned on bin with a full Bayesian network through a full Bayesian network (described above in Section 1.5.4.1: Bayesian Methods and in more detail in Chapter 2).

#### 1.6.2.4 Transitive (T)

The Transitive (T) module is based on the principle that if Protein A interacts with Protein B and Protein B interacts with Protein C, then Protein A is more likely than not to also interact with Protein C. As evidence, the Transitive module considers the topology of the interaction network predicted by the Expression, Orthology and Combined modules to calculate a neighbourhood topology, or transitive, score by assessing the shared interactions between the interaction partners of the two query proteins (Scott & Barton, 2007). If the individual proteins in the pair share one or more interacting partners, a neighbourhood topology score is computed by dividing the sum of the EOC scores between the shared interactions by the sum of the differences in edges not shared by the two proteins (Equation 1.17).

$$T = \frac{\sum_{e \in E_c} s_e}{1 + \left|E_i \setminus E_c\right| + \left|E_j \setminus E_c\right|}$$

**Equation 1.17: Transitive Neighbourhood Topology Score.** $E_i$ is the set of edge for protein *i*, $E_j$ is the set of edges for protein *j*, $E_c$ is the set of common edges between proteins i and j, se is the likelihood ratio for each common edge between proteins *i* and *j*, $E_i \setminus E_c$ is the difference of edges between $E_i$ and $E_c$ and $E_j \setminus E_c$ is the difference of edges between $E_j$ and $E_c$.

#### 1.6.2.5 Cluster (M)

The Cluster (M) module was introduced in v. 2.0 (McDowall, 2011) as a second option for network analysis. First, pairs included in the preliminary interaction network are grouped into clusters using the Markov Clustering (MCL) algorithm (Enright *et al*., 2002). Briefly, the MCL algorithm is an unsupervised clustering method that simulates random walks connecting pairs of entities within a graph or network (i.e. 'edges' connecting 'nodes') to group the network into subnetworks. After between three and ten iterations of alternating 'expansion', where clusters attempt to acquire new nodes by taking longer 'walks', and 'inflation', where clusters increase their connections within the cluster to eliminate connections between other clusters, the network is divided into small groups. While other clustering algorithms exist (i.e. hierarchical and *k*-means), the MCL algorithm was selected for PIPs as it does not rely on any prior knowledge about the clusters or the network (McDowall, 2011).

The performance of clustering algorithms can be assessed in two ways: accuracy, or how correctly members of the same complex are grouped into the same cluster, and separation, or how well different complexes are split from each other (Brohée & van Helden, 2006). As both attributes measure slightly different performances, the ideal algorithm would score highly in both by successfully separating distinct complexes from each other within a cluster while still keeping members of that complex grouped together.

In PIPs, clusters generated by the MCL algorithm are scored as follows (McDowall, 2011): first, all possible interactions in a cluster are assigned, if they have a protein pair

in the training set, the EOC score for them and the partner, or a value of 1. Next, the number of possible interactions in the cluster is calculated as in Equation 1.18.

$$N_x = \frac{n(n-1)}{2} \qquad \text{where} \qquad n = |C_x|$$

**Equation 1.18: Calculation of the number of edges in a complete cluster.** $N_x$ is the number of possible interactions in a cluster and $n$ is equal to the a given cluster $C_x$.

The total cluster score (Equation 1.19) is the sum of the $LR_{EOC}$ values for each pair in the cluster divided by the total number of possible interactions.

$$C_{score} = \frac{\sum_{i \in I_t} S_i}{N_x}$$

**Equation 1.19: Calculation of the cluster score.** The cluster score ($C_{score}$) is calculated by dividing the sum of the $LR_{EOC}$ values for each protein pair in the cluster divided by the number of possible interactions ($N_x$) within the cluster. If $i$ is an element of $I_t$ (all pairs in the positive and negative training sets), $S_i = LR_{EOC}$, otherwise $S_i = 1$.

The cluster score is dependent upon the number of strong scoring interactions within the cluster; therefore, a large cluster with a few high scoring interactions will not necessarily score better than a small cluster with the same number of high scoring interactions.

## 1.7 Scope of This Thesis

At v. 2.0, both the EOCT and EOCM methods in the PIPs prediction framework are capable of accurately predicting interactions; however, there is ample opportunity for the continued development and practical application of the predictor. Chapter 2 describes the development of PIPs v. 3.0, which includes a wide-range data update, minor coding adjustments and the introduction of a new module (the TransMCL module) that combines the Transitive and Cluster module into one and its associated new prediction method (EOCZ). Chapter 3 details the development of a new framework for PIPs, PIP'NN, that incorporates a neural network in place of the naïve Bayesian network as an alternate method for prediction. In Chapter 4, the predictive capability of PIPs v. 3.0 and PIP'NN are compared both to each other and to other currently available human protein-protein interaction prediction tools. Chapter 5 addresses two different collaborations in which PIPs and PIP'NN were implemented to identify potential interactions in the DNA repair system and as an additional stage of filtering in the SILAC mass spectrometry protocol. In Chapter 6, the new PIPs webserver, which includes an updated front end and backend and includes predictions from v. 3.0, is introduced. Finally, Chapter 7 discusses conclusions from this work and future directions for PIPs and PIP'NN.

# Chapter 2

## PIPs v. 3.0: A New Version of the PIPs Predictor

## Preface

First, this chapter describes the updates to the data within the PIPs database and gives a brief overview of the methodology and any minor changes to the individual modules. It then details the development of a new TransMCL module and the performance of the new predictor in comparison to the previous PIPs predictors.

# 2.1 Introduction

## 2.1.1 Updates to the PIPs Data and Database

With the last major update to the data included in PIPs between 2007 and 2009, the data incorporated both within the individual modules and in the positive interaction set is now several years out-of-date. To allow PIPs to remain current with its relevant data, the set of proteins considered by PIPs, positive dataset and data associated with the individual modules were all updated.

## 2.1.2 Development of the TransMCL (Z) Module

As detailed in Chapter 1.6.1: The PIPs Framework, v. 2.0 of PIPs contains two network analysis modules, the Transitive and Cluster modules, that both consider the same source of evidence: the preliminary interaction network predicted by the Expression, Orthology and Combined modules (McDowall, 2011). Because of this shared evidence, there are two PIPs methods: the EOCT method incorporating the Expression (E), Orthology (O), Combined (C) and Transitive (T) modules and the EOCM method including the Expression, Orthology, Combined and Cluster (M) modules. While the EOCT method predicts known positive and negative interactions with greater accuracy than the EOCM method, inclusion of the EOCM method predicts a distinct set of interactions not predicted by the EOCT method. As a result, the methods are run in tandem to maximise the total number and coverage of predictions (McDowall, 2011).

A full Bayesian network is already included in the PIPs predictor within the Combined module, depicted schematically in Figure 2.1 and described below.

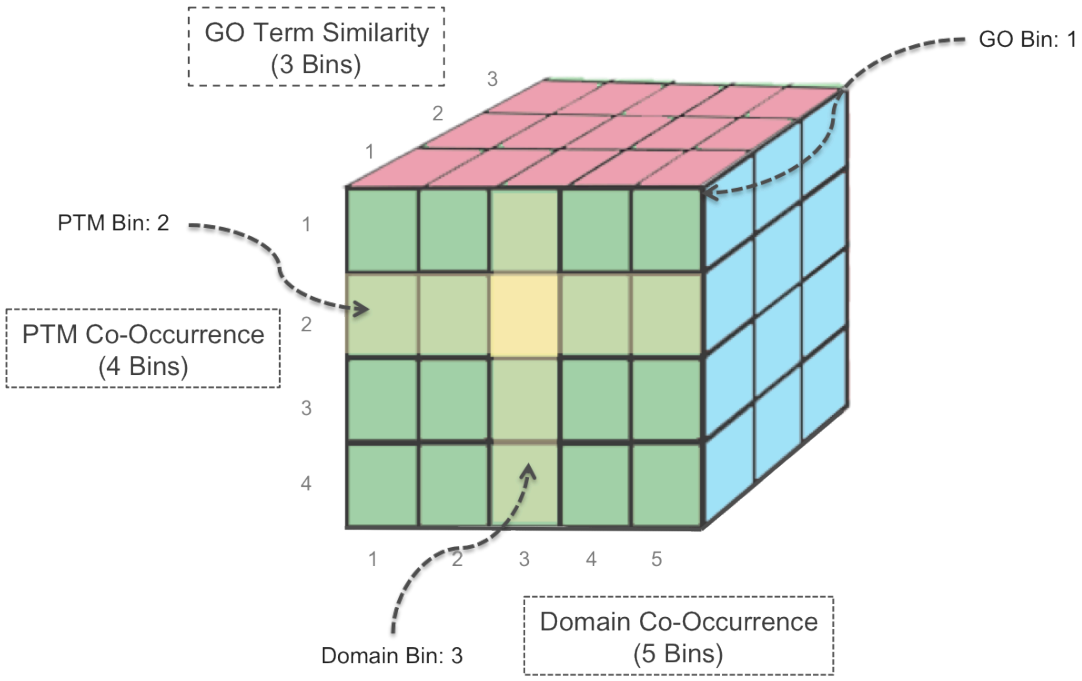


**Figure 2.1: Schematic Diagram of the Allocation of Bins for the Combined Module.** An example of how the correct bin is assigned to a protein pair in a full Bayesian network is shown for the Combined module. For the Combined module, three features are considered, co-occurrence of domains, post-translational modifications and GO terms, which have five, four and three bins, respectively. As the module requires the scores from all three features to calculate a likelihood score, the bins must be grouped together, shown graphically in the diagram as a three-dimensional box split into 3 x 4 x 5 (60) smaller boxes. Each small box represents a possible combination of bins from each feature that the protein pair could be assigned. In this example, the pair has been assigned bin three for the domain feature, bin two for the PTM feature and bin one for the GO terms feature; therefore, it should be assigned in the bin distinguished by the combination of these three bins (shown in yellow).

The Combined module considers three features that are each trained separately. The score for each feature is assigned a bin; for the domain co-occurrence feature, there are five bins, for the post-translational modifications (PTM) feature, there are four bins, and

for the GO terms feature, there are three bins. However, unlike in a naïve Bayesian network, where a likelihood ratio would be calculated for the bin of each feature independently, the bins in the Combined module are grouped together into a 3 x 4 x 5 three-dimensional matrix such that there are 60 possible bins for allocation. In the example in Figure 2.1, the GO term score has been assigned bin one, the PTM score has been assigned bin two and the domain score has been assigned bin three. The correct bin, therefore, corresponds to the small box in the larger three-dimensional box that is in position one in the GO term axis (the front face of the cube), position two in the PTM axis and position three in the domain axis.

With this same approach, it is possible to combine the results from the Transitive module with the results from the Cluster module into one new module. Rather than requiring a three-dimensional matrix for score binning, the new module would include master bins associated with two individual bins: one for the Transitive portion and one for the Cluster portion of the module. For example, if a protein pair had a transitive score placing it in transitive bin one and a cluster score placing it in cluster bin two, its final new bin would be the bin one-two, which would be assigned a likelihood ratio based on the number of positive and negative pairs associated with that combination of bins.

Ideally, combining the methods of analysing the initial predicted network underlying the Transitive and Cluster modules will have a cumulative effect of increasing the number of predicted interactions while improving the accuracy. If successful, the new module will be able to take the place of both the Transitive and Cluster modules, allowing for one prediction method.

# 2.2 Methods

## 2.2.1 Updates to the Protein Dataset

The proteins included in the PIPs database were last updated in 2009 with the human proteins from EBI's International Protein Index (IPI) database, which includes an aggregation of proteins from the UniProtKB/TrEMBL (UniProt Consortium, 2012), Ensembl (Flicek *et al.*, 2012), Unigene (Mayer, 2008), Vega (Wilming *et al.*, 2008), RefSeq (Pruitt *et al.*, 2012) and H-InvDB databases (Yamasaki *et al.*, 2010). To accommodate for any new proteins that have been discovered over the past three years, any new, reviewed, additions to the UniProtKB/Swiss-Prot database between the dates of the date of the last protein set update in PIPs (25 March 2009) until the then-current date (12 May 2012) were downloaded.

To ensure that the new proteins were not already included in the database, the new identifiers and amino acid sequences were searched against the current PIPs protein entries and any matches eliminated. This update resulted in 299 new proteins in the PIPs database.

Since September 2011, the IPI database has been deprecated. To accommodate this change, the main identifiers for the proteins in PIPs were updated to match their identifiers in each of their original, non-IPI source databases. The IPI cross-reference file downloaded from the IPI website (May 2012) was used to map each IPI identifier in the current PIPs database with its identifier from the following sources in preferential order:

UniProtKB/Swiss-Prot Accession

UniProtKB/trEMBL Accession

Ensembl

Vega

Unigene

RefSeq

H-InvDB.

Table 2.1, below, details the breakdown of the IPI to cross-reference mapping and the new contents of the protein dataset within PIPs.

| Source | Number of Entries |
|---|---|
| UniProtKB | 50,282 |
| trEMBL | 11,823 |
| Ensembl | 1471 |
| Vega | 8555 |
| Unigene | 2528 |
| RefSeq | 1649 |
| HInv | 1704 |

**Table 2.1: Details of IPI Cross-Reference Update to the Proteins in the PIPs Database.** The number of proteins in the PIPs database from the IPI database with their main identifier mapped to one of the seven original sources in the IPI is provided above. Identification references were parsed from the cross-reference file downloaded from the IPI website.

While within the PIPs database itself, proteins are identified by numerical IDs (i.e. 1 through 76,799 for the human proteins), both updating the database with data for each of the sources of evidence and practically utilising the database to search for predicted

interactions requires data to be accessible by a range of external identifiers. To allow for this searching capability, the PIPs database also contains a cross-reference table (Other_Accession) that includes any additional identifiers associated with each protein entry. Therefore, to complete the update of the main protein identifiers, all IPI identifiers and any additional references present in the IPI cross-reference file were added to the Other_Accession table. Additionally, the identifiers for the newly added proteins were mapped through the EBI's PICR cross-referencing tool and added to the database.

As the PIPs database includes isoform variants of several proteins, it was necessary to assign one of the variants as the primary protein entry. Where a clear, main entry for a protein with one or more isoforms was not available, the first variant was taken as the main entry. For all variants, the other isoform variants were added to the Other_Accession table to maximise protein mapping.

Following both updates, there are now 76,799 unique protein entries within the PIPs database. This number includes isoformic variants listed as separate entries.

### 2.2.2 Reconstruction of the Positive and Negative Datasets

The datasets included in the current version of PIPs are described in detail in McDowall, 2011. The positive dataset was derived from the Human Protein Resource Database (HPRD) (Prasad, n.d.), a manually curated database over ~39,200 (as of August 2012) binary protein interactions verified by experimental techniques. As there is no informative database for negative interactions, the negative dataset was assembled

by randomly selecting protein pairs and filtering out any that do interact according to the HPRD (Prasad, n.d.), IntAct (Kerrien *et al.*, 2012), BioGRID (Stark *et al.*, 2011), DIP (Salwinski *et al.*, 2004) and OPHID (Brown & Jurisica, 2007) databases.

To update the positive dataset (derived from the 2007 release of the HPRD in PIPs v. 2.0), the most recent HPRD update (Release 9) was downloaded from the HPRD website and replaced the existing positive datasets in full with 38,995 interactions. The negative dataset was then reassembled, as describe above, by selecting 100x the number of positive interactions and filtering any protein pairs recorded as interacting in any of the above sources.

For training and testing, the datasets were split randomly into six groups, five for five-fold cross-validation plus an additional group as a blind test set. For ease and time conservation, the datasets were split permanently once, and after cross-validation training, the Expression, Orthology and Combined scores for pairs were pre-calculated and stored in the PIPs database, allowing these values to be accessed during development of the network analysis modules without repeated re-computation.

### 2.2.3 Prior Odds Ratio

The prior odds ratio, the estimate of how likely an interaction is to occur not by chance, has been kept as 1/1000 as described in McDowall, 2011.

### 2.2.4 Interactome Database

All evidence of and information regarding interactions are included in the local Interactome database. Currently, the database is 636 GB in size and stored in MySQL version 5.0.45.

### 2.2.5 Modifications to the PIPs v. 2.0 Modules

PIPs v. 2.0 contains three main modules (the Expression, Orthology and Combined modules) that incorporate six different pieces of evidence derived from outside sources. Details of the data included in each module, a brief summary of the module's methodology and any updates made from the original version of PIPs are outlined below. For a complete description of previous developments in each of the modules, refer to Scott and Barton, 2007 and McDowall, 2011 and Chapter 1.6.2: PIPs Modules.

#### 2.2.5.1 Expression

The acquisition and development of the expression datasets is described in McDowall, 2011. Briefly, intensity values were taken from probes for 18,334 proteins from the E-GEOD-7307 microarray (Roth *et al*, 2006). The Pearson's Correlation Coefficient, which ranges from -1.0 to 1.0, for each pair of proteins was calculated to quantify the likelihood that the two proteins are expressed simultaneously (McDowall, 2011).

Gathering of the original data incorporated into the Expression module was a lengthy and in depth process, and repeating the entire methodology with a slightly more recent dataset was not thought to produce significantly different scores. Therefore, the expression scores dataset was left unchanged from PIPs v. 2.0.

To assign a final Expression module score, each pair of proteins is assigned one of 20 bins, each of which spans a 0.1 score range window, based on its calculated Pearson's Correlation Coefficient.

#### 2.2.5.2 Orthology

The orthologue mappings for the PIPs proteins are derived from the InParanoid (Ostlund *et al.*, 2010), BIND (Bader, Betel & Hogue, 2003), DIP (Salwinski *et al.*, 2004) and GRID databases (Stark *et al.*, 2011). Scores are provided by InParanoid and reflect how closely related the orthologous proteins are relative to the two most closely related 'seed' orthologues (see Chapter 1.6.1.1.2: Orthology).

The new InParanoid dataset was downloaded on 31 October 2011, any entries not already included in the current PIPs database were added and any necessary changes to existing data made. Currently, the InParanoid data in the PIPs database reflects the Version 7 release downloaded from the InParanoid website (http://inparanoid.sbc.su.se).

The Orthology module follows a multi-step process of analysing interactions between orthologues of proteins in PIPs. During the first step, all yeast, worm and fly orthologues included in the PIPs database for each of the proteins in the pair in question are attained. Next, the database is queried for any known interactions between the orthologues of the two proteins of interest. Depending on how many interactions are known between the orthologous proteins, the pairs are assigned one of five bins, where bin zero represents no known interactions, bin one represents a known interaction in one of the species, bin two represents known interactions in two of the species and so on.

#### 2.2.5.3 Combined

The Combined module incorporates evidence about post-translational modifications, the co-occurrence of InterPro domains and Gene Ontology (GO) terms (replacing subcellular localisation from v. 1.0) to assign one score representing the three features (see Chapter 1.6.2.3: Combined Module). This score is calculated through a full Bayesian network (as described in more detail in Section 2.1.2: Development of the TransMCL Module, above) with a 3 x 4 x 5 bin matrix leading to 60 possible likelihood ratios.

##### 2.2.5.3.1 Domains

To update the domain portion of the Combined module, PFAM domain associations were updated in October 2011 with the Version 25 release from the PFAM website (http://pfam.sanger.ac.uk) and any new entries added and existing entries modified as necessary. For the new proteins added to the PIPs database, InterPro domains, which include PFAM domains, were downloaded along with the original protein data from the UniProtKB database.

To assign a bin for each protein pair in the domain portion of the Combined module, both InterPro motifs and domains and PFAM domains are considered. First, a Chi-squared value was calculated for each InterPro domain-domain and motif-motif pairing based on how many times the pairing is seen among all known interacting proteins. If the query protein pair contains one or more of the motif pairings, it is assigned the highest Chi-squared score from the pairings and one of five bins covering an increasing range of scores. Second, if the query pair contains any of a set of PFAM domain pairs

seen in interacting proteins from structural studies, they re automatically assigned the bin for the highest Chi-square scores.

#### 2.2.5.3.2 GO Terms

Where possible, the proteins in the PIPs database were matched to their associated GO terms downloaded from the Gene Ontology Database (http://www.geneontology.org). Protein-GO term associations were updated in December 2011 with the human GO gene association file (CVS version 1.220, 13 December 2011 release). After calculation of the semantic similarity between two terms (described in full detail in Chapter 1.6.2.3: GO Term Similarity), GO scores were assigned to one of three bins covering the range of possible scores.

#### 2.2.5.3.3 Post-Translational Modifications

Post-translational modification (PTM) information was downloaded from the UniProtKB website and updated in November 2011. PTMs for the new PIPs proteins were downloaded in May 2012 from the UniProtKB website.

A PTM score for each pair of proteins is calculated by dividing the number of times that a specific PTM pairing is seen in the total set of PTMs occurring across an interaction dataset by the number of times each PTM is seen across the interaction set on its own. Each score was then grouped into one of four bins covering the range of possible scores.

#### 2.2.5.4 Transitive

The Transitive module requires the input of the network of interactions predicted by the Expression, Orthology and Combined modules to calculate a neighbourhood topology score, as described in Scott and Barton, 2007 and McDowall, 2011. To generate this network, the likelihood ratios for the Expression, Orthology and Combined modules on their own are calculated for each pairing of proteins in the selected dataset and are multiplied together to give the preliminary $LR_{EOC}$ score. Predictions were then filtered to remove all pairs with $LR_{EOC}$ values less than a cut-off threshold of 10. The remaining pairs then form the preliminary predicted interaction network.

Protein pairs are assigned a transitive score (as described in Chapter 1.6.2.4: Transitive Module) and then are assigned one of four bins corresponding to increasing transitive scores.

#### 2.2.5.5 Cluster

The preliminary interaction network considered by the Cluster Module was generated as described above with the exception that instead of a cut-off of 10.0, a less stringent threshold of 5.0 was chosen (McDowall, 2011). Interacting proteins were then grouped into clusters by implementing the Markov Clustering (MCL) algorithm (Enright, Van Dongen & Ouzounis, 2002) as described in McDowall, 2011 (see Chapter 1.6.2.5: Cluster Module for more detail). Protein pairs were assigned a cluster score based on if the two proteins are grouped in the same cluster and how many known interacting pairs of proteins are included within that cluster. The pair is then assigned one of five bins covering an increasing range of scores.

To increase the efficiency of the Cluster module, the code was rewritten to store the cluster groupings and scores for each cluster rather than recalculating them multiple times through training and testing. This modification drastically decreased the module's training runtime from over 24 hours to under one hour.

## 2.2.6 The TransMCL Module (Z)

### 2.2.6.1 TransMCL bins

The Transitive and Cluster modules were combined through a full Bayesian network similar to that described for the Combined module (see Section 2.1.2: Development of the TransMCL Module) to form a new module, the TransMCL (Z) module. While most of the code for the Transitive and Cluster modules was preserved, the minor changes made to increase the efficiency of the Cluster module were also implemented in the TransMCL module. Although the lower cut-off for the Transitive module is 10.0, the threshold for the Cluster module, 5.0, was selected as the lower cut-off for $LR_{EOC}$ scores for an interaction pair to be included in the initial network of interactions for the TransMCL module. This lower cut-off ensured that the proteins incorporated into the network supplied as evidence to the Cluster module but not the Transitive module were not excluded from the TransMCL module.

During the first stage of the TransMCL module, the transitive score for the network of interactions produced by the EOC modules was calculated for each protein pair, and each pair was assigned one of five bins covering the range of potential scores. After testing several different score ranges for each of the bins, bin groupings were kept the

same as for the Transitive module, except the upper limit for bin three and lower limit for bin four were changed from 1600 to 1000. Next, a cluster score was calculated for each pair based on the same EOC network, and the pair was assigned to one of five bins covering the range of potential cluster scores. Again, after testing several different score ranges, bin groupings were kept the same as in the Cluster module on its own. The two bins were then combined into a matrix of 5 x 5 total bins covering every possible combination of transitive and cluster scores to give a total of 25 bins for the TransMCL module. Table 2.2 describes the bin dimensions implemented in the final version of the module.

| | Scores Included | | Scores Included |
|---|---|---|---|
| **Transitive 0** | < 25 | **Cluster 0** | No Cluster Score |
| **Transitive 1** | ≥ 25 or < 100 | **Cluster 1** | ≤ 50 |
| **Transitive 2** | ≥ 100 or < 400 | **Cluster 2** | > 50 or ≤ 200 |
| **Transitive 3** | ≥ 400 > 1000 | **Cluster 3** | > 200 or ≤ 500 |
| **Transitive 4** | ≥ 1000 | **Cluster 4** | > 500 or ≤ 1000 |

**Table 2.2: Bin groupings for the TransMCL Module.** Upper and lower thresholds for each of the five Transitive and five Cluster bins for the TransMCL module are provided. Transitive bins 3 and 4 were changed from having an upper limit and lower limit of 1600 to 1000, respectively. The number of Cluster bins was reduced from 6 to 5 by altering the range of coverage for each of the bins.

As in the other modules, a likelihood ratio was then calculated for each pair of bins in the matrix by counting the total numbers of positive Transitive and Cluster interactions,

negative Transitive and Cluster interactions, and dividing the values by the total number of positive and negative interactions.

### 2.2.7 Retraining PIPs

After updating the protein and individual module data and the positive dataset, the entire PIPs predictor was retrained and retested. Figure 2.2 shows the three stages of the PIPs training and prediction pipeline. Further details of the cross-validation, full training and full prediction stages are provided in the sections below.

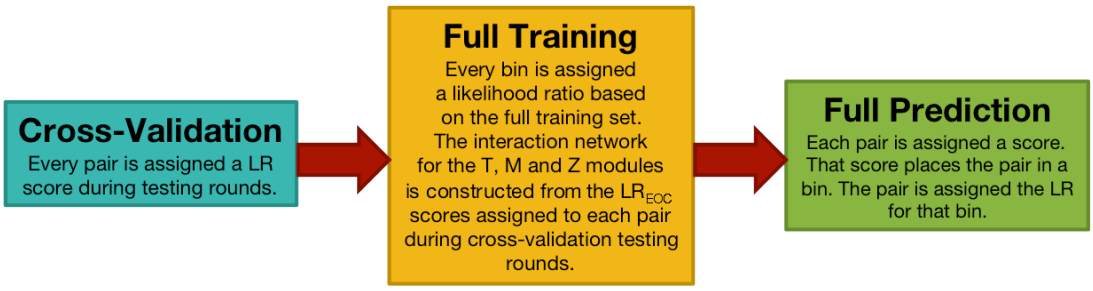


**Figure 2.2: Schematic of the PIPs training and prediction pipeline.** Training, testing and prediction in PIPs follows three stages. First, the predictor is trained with five-fold cross-validation, during which each pair is assigned a likelihood ratio for each module when it is part of the test set. Next, the predictor is retrained on the full training dataset. This stage assigns a final training likelihood ratio to each bin in each module. For the Transitive, Cluster and TransMCL modules, the interaction network supplied as evidence is constructed from the Expression, Orthology and Combined likelihood ratios assigned to each pair during the cross-validation testing rounds. Finally, the full set of predictions is generated by going through each module and calculating the appropriate evidence score for the pair that allocates it to a bin. The pair then assumes the likelihood ratio assigned to the bin.

#### 2.2.7.1 Cross-Validation and Full Training

First, each of the Expression, Orthology and Combined modules was independently trained with five-fold cross-validation in the same manner as the original predictor as described in Scott and Barton, 2007 and McDowall, 2011. Briefly, the positive and negative datasets were split into five identically sized groups. The predictor was then

trained by selecting four of the datasets for training, during which a likelihood score was calculated for each bin based on the positive and negative interactions in these four datasets. The fifth dataset was then run as a test, where its protein pairs were assigned a bin and given the likelihood score of the bin that was allocated during the training. The training and testing process was repeated five times rotating the training and testing datasets such that each set was used for testing once. After cross-validation was complete, each protein pair had been in the test set and had been assigned a likelihood ratio.

Each of the modules was then trained on the full positive and negative training datasets (i.e. subsets one through five) to assign a final likelihood ratio to each bin. The preliminary interaction network supplied to the Transitive, Cluster and TransMCL modules as evidence was constructed based on the Expression, Orthology and Combined likelihood ratios assigned to each pair during the testing rounds of cross-validation.

#### 2.2.7.2 Generation of the Full Prediction Set

To ensure that predictions reflect a true assessment of the likelihood of interaction and are not based purely on a lack of evidence, the set of PIPs predictions contains all possible pairing of proteins within the PIPs protein dataset for which there is at least one source of evidence available. To generate the PIPs prediction set, predictions were first made for the Expression, Orthology and Combined modules separately by assessing the available evidence for the pair and scoring it with the likelihood ratio, calculated during full training of the module, that was associated with its assigned bin. Once the set of

Expression, Orthology and Combined predictions were made, the $LR_{EOC}$ score for each pair was calculated by multiplying the three individual likelihood ratios together.

For the Transitive module, the set of all possible protein pairs was filtered to include only those with $LR_{EOC}$ scores above 10.0. This filtering resulted in a preliminary interaction network with 1,190,825 protein pairs. Each pair in the set of all possible pairs was assigned a transitive score based on this network, the bin associated with that score and finally, the likelihood ratio for that bin.

For the Cluster module, the set of all possible protein pairs was filtered to include only those with $LR_{EOC}$ values above 5.0. The resulting network, which contained 3,348,424 protein pairs, was then grouped into clusters by the Markov Clustering (MCL) algorithm (Enright, Van Dongen & Ouzounis, 2002). Protein pairs were then assigned a cluster score based on these clusters, the bin associated with that score and the likelihood ratio for that bin.

The TransMCL module considers the interaction network provided as evidence to the Cluster module (i.e. including 3,348,424 protein pairs with $LR_{EOC}$ scores above 5.0). Pairs were assigned a transitive score based on this network and a cluster score based on the same cluster groupings as in the Cluster module, the individual bins associated with each score and the likelihood ratio for the transitive-cluster combination of bins.

Once the individual scores had been computed for each of the modules, the likelihood ratios for each pair were calculated by multiplying $LR_{EOC}$ score by the likelihood ratio for the Transitive module (for the EOCT predictor), the Cluster module (for the EOCM

predictor) or the TransMCL (for the EOCZ predictor). Finally, this $LR_{EOCT}$, $LR_{EOCM}$ or $LR_{EOCZ}$ was multiplied by the 1/1000 prior odds ratio to give the final prediction score. The final prediction score represents how many more time likely than not the protein pair is to interact.

### 2.2.8 Validation of Accuracy

The accuracy of the EOCT, EOCM and EOCZ predictors was quantified through Receiver Operator Characteristic (ROC) plots by calculating the sensitivity (the number of true positives) and the specificity (the number of false positives) of predictions. ROC100 curves, which rank positive and negative predictions in decreasing order and then compare the number of highest scoring true positive results predicted before the 100th highest scoring false positive result is returned, were plotted at two main stages: 1) after cross-validation training to compare how well each of the three predictors performed, on average, on the test set not included in training and 2) after attaining the final EOCT, EOCM and EOCZ scores for protein pairs in a blind test set of 5000 positive and 5000 negative randomly chosen pairs not included in the five training data subsets.

To compare the prediction accuracy of the three versions of PIPs, a second blind test set with 2588 positives and 2588 negatives was chosen by randomly selecting pairs from the blind test set of PIPs (v. 3.0) that were also included in PIPs v. 1.0 and 2.0.

To analyse the numbers of predictions made by each of the predictors, the lower cut-off for significant interactions was set at 1.0 after multiplication of the $LR_{EOCT}$, $LR_{EOCM}$ or $LR_{EOCZ}$ score by the 1/1000 prior odds ratio.

# 2.3 Results

## 2.3.1 Prediction Accuracy of the EOCT, EOCM and EOCZ Predictors during Cross-Validation Testing

Although it was hoped that merging the Transitive and Cluster modules into one, cumulative module would increase the number of predictions and, more importantly, the accuracy of PIPs, the full EOCZ predictor failed in both regards. Figure 2.3, below, shows the ROC100 curves for each of the three PIPs predictors comparing the average number of false positive results before the 100th true positive result across the five test sets from cross-validation. Of the three methods, the EOCT predictor still predicts a larger number of pairs in the positive set with high scores (609) than the EOCM (463) and EOCZ (558) methods before 100th highest-scoring negative. However, looking further at the highest scoring predictions that fall in the left-hand lower corner of the plot (i.e. the first true and false positives predicted) shows that the EOCZ predictor starts off identifying a greater number of true positives (93) before the first false positive is predicted than the EOCM (43) and EOCT (78) predictors. Despite this initial better performance of the EOCZ method, as the numbers of false positive results increase, the EOCZ method performs in between that of the EOCT and EOCM methods overall. This pattern suggests that incorporation of the Transitive module on its own is the most effective means of correctly identifying interacting pairs of proteins.

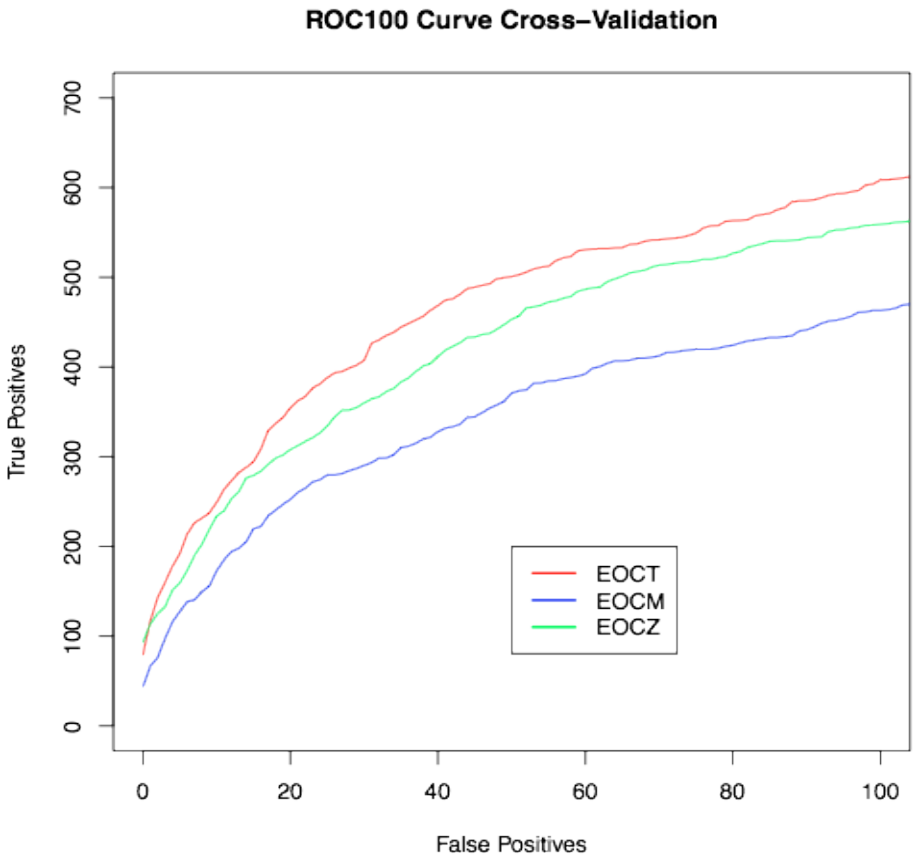


**Figure 2.3: ROC100 plot of the average true positive and false positive predictions during cross-validation testing.** The ROC100 curves plotted for the average number of true positives predicted with the highest likelihood ratios before the first 100 false positives are predicted for the EOCT (red), EOCM (blue) and EOCZ (green) predictors are shown above. To construct the plot, the highest scoring predictions for each of the methods were ranked in descending order and grouped as either a true positive (for those in the positive dataset) or a false positive (for those in the negative dataset). True and false positive values are represented as the average of the absolute count for each test set during the five rounds of cross-validation.

### 2.3.2 Prediction Accuracy of the EOCT, EOCM and EOCZ Predictors in a Blind Test

To confirm the difference in accuracy between the three PIPs methods, the final likelihood ratios for the a blind test set of protein pairs containing 5000 positive and 5000 negative examples were compared through a ROC100 plot, shown in Figure 2.4, below.

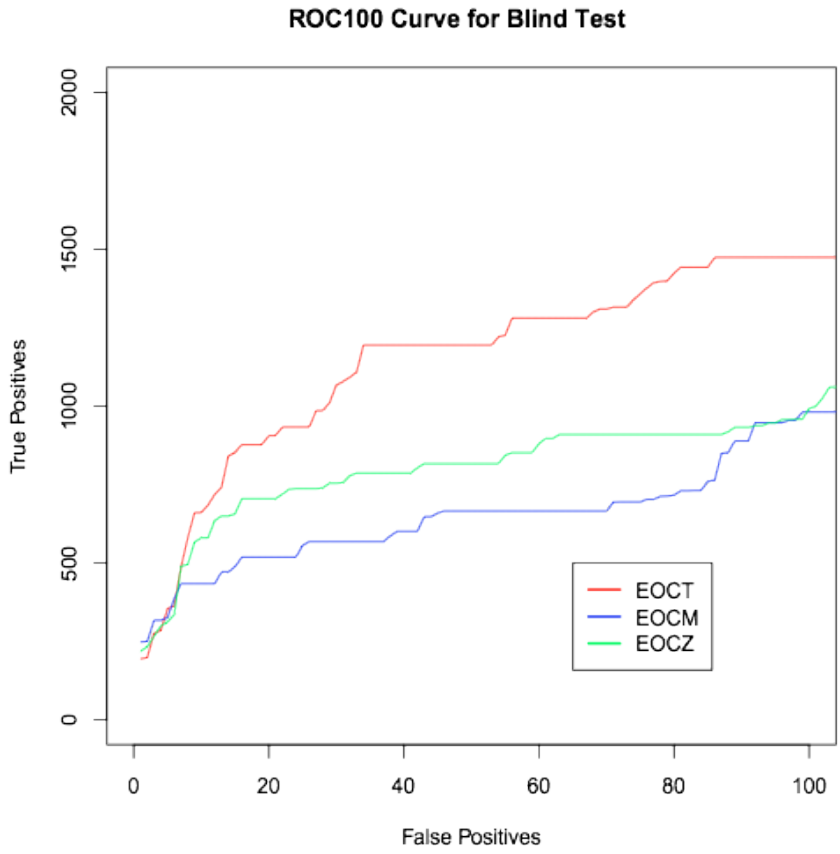


**Figure 2.4: ROC Plot comparing the EOCT, EOCM and EOCZ prediction methods.** The number of true positive results attained before the first 100 false positive results from a blind test with 5000 positives and 5000 negatives are plotted as a ROC100 curve for the EOCT (red), EOCM (blue) and EOCZ (green) methods. Pairs for the test were selected at random from the full blind dataset containing protein pairs not seen by the predictor during training.

The higher number of true positive predictions before the 100th false positive result for the EOCT method (1474, compared to 981 for the EOCM method and 997 for the EOCZ method) follows the same pattern as seen during cross-validation (Figure 2.3, above), confirming that incorporation of the Transitive module is more effective at correctly identifying known positive and known negative interactions than either the Cluster or TransMCL modules.

To investigate if there were any discernible patterns in the pairs of proteins in the negative blind test set that were assigned the highest likelihood ratios, the top 12 highest scoring false positive predictions from the EOCT, EOCM and EOCZ predictors were compared. Table 2.3, below, shows the protein pair and the three prediction likelihood ratios for the four overlapping pairs that were in the highest scoring sets for all of the methods.

Interestingly, of the protein pairs in these high scoring false positive sets, the most common type of interaction involved two zinc finger proteins (5/20 from EOCT PIPs, 4/20 from EOCM PIPs and 7/20 from EOCZ PIPs). Looking closer at the module contribution for each of these zinc finger-containing pairs and the predictions shared by all three predictors showed that each was driven by a moderately high likelihood ratio for the Orthology and the Combined modules. The Combined module scores are slightly surprising; while zinc finger-zinc finger protein interactions have been described (Imanishi *et al.*, 2010), the domain is typically involved in mediating protein-DNA or protein-RNA interactions. However, with the Orthology module as a strong indicator of interaction, it is possible that these predictions are indicative of genuine interactions.

| Protein1 | Protein2 | EOCT Score | EOCM Score | EOCZ Score |
|---|---|---|---|---|
| **H2AFV** (Histone H2A.V) | **HIST1H2AC** (Histone H2A type 1C) | 89091 | 26472 | 407930 |
| **ZNF174** (Zn-finger protein) | **ZSCAN10** (Zn-finger and SCAN domain-containing protein) | 485035 | 144123 | 2220860 |
| **ZNF749** (Zn-finger protein) | **ZNF316** (Zn-finger protein) | 65717 | 17369 | 394771 |
| **ZNF7** (Zn-finger protein) | **ZNF8** (Zn-finger protein) | 24807 | 6556 | 101027 |

**Table 2.3: Selected highest scoring false positive predictions shared across the EOCT, EOCM and EOCZ predictors in the blind test set.** Likelihood ratios are given prior to adjustment for the 1/1000 prior odds ratio. Of the top 20 highest scoring false positive predictions for each of the three predictors, the four above were the only ones shared across the set.

Several other interactions within these false positive sets are also of interest; in particular, there are two histone-histone interactions included in the top scoring false positives from the EOCT predictor, one of which (H2AFV and HIST1H2AC) is also assigned a high score from the EOCM and EOCZ predictors. Again, assessing the module contribution for this prediction revealed that, in addition to high Transitive, Cluster and TransMCL module likelihood ratios, the Orthology module ratio is high. Coupled with the knowledge that these two proteins share similar functions and are involved in similar processes, this recognition of known interactions between the two proteins in other species suggests that this might be an interaction that is not yet recorded. Finally, another pair of interest predicted by the EOCM method is between ACTG1, a cytoplasmic variant of actin, and PHACTR4, one isoform of the phosphatase and actin co-regulator. Again, this high score is dependent largely upon a high

likelihood ratio in the Orthology module; however, it also seems reasonable that these two proteins might interact based on function alone.

### 2.3.3 Comparison of the Transitive, Cluster and TransMCL Modules

The poorer performances of the Cluster and TransMCL modules in comparison with the Transitive module warranted further investigation into how protein pairs were being assessed, binned and scored in the modules. First, to analyse how the TransMCL module was handling each of its Transitive and Cluster module components, the number of positives and negatives assigned to each bin were extracted from training of the full predictor on the five training datasets. Table 2.4, below, shows a two-dimensional table with the absolute counts for positives (top number in each cell) and negatives (bottom number in each cell), with the Transitive portion of each bin represented horizontally and the Cluster portion represented vertically. Looking at the proportion of positives to negatives for each Transitive-Cluster pair of bins gives an indication of how the method is handling pairs with different combinations of scores for the two components.

The low-number combination bins, particularly those with either the Transitive bin as 1 or the Cluster bin as 1, include pairs that score very low for one of the components, making binning and scoring reliant upon the second component. Likewise, the high-number combination bins include pairs that score highly for one of the components. For example, if a pair A-B scores very low for the Cluster component but very high for the Transitive component, it would be placed in the combination bin 5-1. Likewise, if a pair C-D scores very low for the Transitive component but very high for the Cluster

component, it would be placed in the combination bin 1-5. Ideally, these lopsided bins would be the primary way that it could merge predictions for both components.

| Trans MCL | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 25748<br>3240075<br>(0.841) | 501<br>126<br>(420.729) | 311<br>67<br>(491.158) | 161<br>19<br>(896.619) | 124<br>14<br>(937.194) |
| 2 | 1971<br>7476<br>(27.897) | 142<br>49<br>(306.639) | 120<br>24<br>(529.061) | 59<br>8<br>(780.365) | 18<br>2<br>(952.310) |
| 3 | 553<br>842<br>(69.494) | 45<br>6<br>(793.591) | 74<br>8<br>(978.763) | 44<br>6<br>(775.956) | 56<br>7<br>(846.497) |
| 4 | 214<br>283<br>(80.013) | 5<br>2<br>(264.530) | 11<br>2<br>(581.967) | 14<br>0<br>(1.403E-10) | 9<br>0<br>(3.207E-10) |
| 5 | 361<br>362<br>(105.520) | 22<br>2<br>(1163.934) | 20<br>5<br>(423.249) | 32<br>0<br>(9.019E-11) | 94<br>1<br>(9946.344) |

**Table 2.4: Number of positives and negatives assigned to each bin during full training of the TransMCL module.** The table above shows the breakdown of positive and negative pairs assigned to each Transitive-Cluster combination bin during full training of the TransMCL module. Each combination bin contained two sub-bins: a Transitive bin (shown in red and corresponding to columns 1-5) and a Cluster bin (shown in blue and corresponding to rows 1-5). Positive counts are shown as the top number in each cell with negative counts as the number underneath and the calculated likelihood ratio in parentheses. Bins in column 1 (light blue) include pairs that have no or a very low transitive score and a cluster score that increases in value as the number of bin increases. Likewise, bins in row 1 (light pink) contains pairs with no or a very low cluster score and a transitive score that increases in value as the bin number increases. Ideally, the proportion of positives:negatives in the 'no transitive' bins (light blue) and the 'no cluster' bins (light pink) should show a higher number of positives to a low number of negatives, indicating that the method used to group pairs in that bin is able to discriminate between scoring positive and negative examples.

Therefore, the combination bins corresponding to low or no transitive scores (i.e. bins 1-[1-5]) and to low or no cluster scores (i.e. bins [1-5]-1) should contain similar high number of positives and low number of negatives. However, the difference between the bins with no transitive score and those with no cluster score is pronounced. While the no-cluster bins include more positives than negative pairs with the positive:negative ratio increasing as the transitive bin increases, the no-transitive bins show the opposite effect. At even the highest no-transitive bin 1-5, which should contain pairs with very high cluster scores suggesting interaction, the number of negatives is still one more than the number of positives assigned to that bin. While the absolute numbers of positives and negatives for each of the no-transitive bins is greater than the numbers assigned to the no-cluster bins, the lack of discrimination between the two example sets suggests that assignment of a high cluster score does little to indicate a positive or negative interaction.

With this in mind, the positive and negative bin counts for the Transitive and Cluster modules on their own, shown in Table 2.5, were also examined and found to share similar patterns of pair distributions as in the TransMCL module.

This distribution confirms that on its own, the Cluster module also performs poorly. However, the difference between the number of pairs that the Cluster and Transitive modules consider with enough evidence to assign a bin (i.e. for the Cluster module, the pair is part of one of the cluster groups and for the Transitive module, the pair has shared interactors that can be assessed), is over three-million pairs, suggesting that the network analysis of the MCL algorithm and pair clustering is much more non-specific than the interaction grouping in the Transitive module. This assignment of binnable

cluster scores to nearly every protein pair could be causing the module to lose accuracy by incorporating proteins into larger clusters that lack true potential for an interaction.

| Bin | Transitive Module | Cluster Module |
|---|---|---|
| 1 | 15793<br>281538<br>(0.897) | 27366<br>3242013<br>(0.893) |
| 2 | 662<br>233<br>(45.41) | 1978<br>6096<br>(34.33) |
| 3 | 535<br>74<br>(115.5) | 354<br>415<br>(90.26) |
| 4 | 317<br>39<br>(129.9) | 121<br>126<br>(107.3) |
| 5 | 332<br>25<br>(212.2) | 370<br>223<br>(101.6) |
| 6 | N/A | 520<br>513<br>(175.6) |

**Table 2.5: Number of positives and negatives assigned each bin in the Transitive and Cluster modules on their own.** Counts are given for positive and negative examples assigned to each of the five Transitive module bins (red column) and to each of the five Cluster module bins (blue column) during full training of both modules on their own. Likelihood ratios calculated for each bin are shown below the counts in parentheses.

Therefore, while the MCL algorithm in the Cluster module is capable of scoring positive interactions highly, its inability to adequately discriminate against negative interactions substantially lowers its prediction reliability in both the EOCM and EOCZ predictors.

### 2.3.4 Comparison of the EOCT, EOCM and EOCZ Final Prediction Sets

In total, final EOCT, EOCM and EOCZ prediction likelihood ratios were calculated for 704,832,309 protein pairs. The breakdown of numbers of pairs with a likelihood of interacting greater than 1.0, shown as a barchart in Figure 2.5, below, reveals an interesting pattern of prediction across the three predictors. First, the EOCT predictor, with 240,663 pairs, predicts about 50% more interactions as the EOCM predictor (162,323) and but less than half of the interactions returned by the EOCZ predictor (579,247). The total number of shared predictions (126,107) represents 78% of the potential overlap that could occur between the methods.

The EOCZ method, which was hoped to increase the coverage of prediction, predicted the majority of interactions predicted by the EOCM (128,295 or 79% of the total EOCM results) and the EOCZ (224,917 or 93% of the total EOCT results) methods. This coverage suggests that the EOCZ module is capable of incorporating both the Transitive and Cluster modules into one method; however, it has done so with a slight compromise in accuracy compared to the EOCT method (see ROC100 curves in Figures 2.2 and 2.3, above). Taken together with the positive and negative binning patterns of Cluster and TransMCL modules (Table 2.4, below, this breakdown confirms that although the EOCZ predictor does handle the Transitive module portion well, its accuracy is compromised by the Cluster portion.

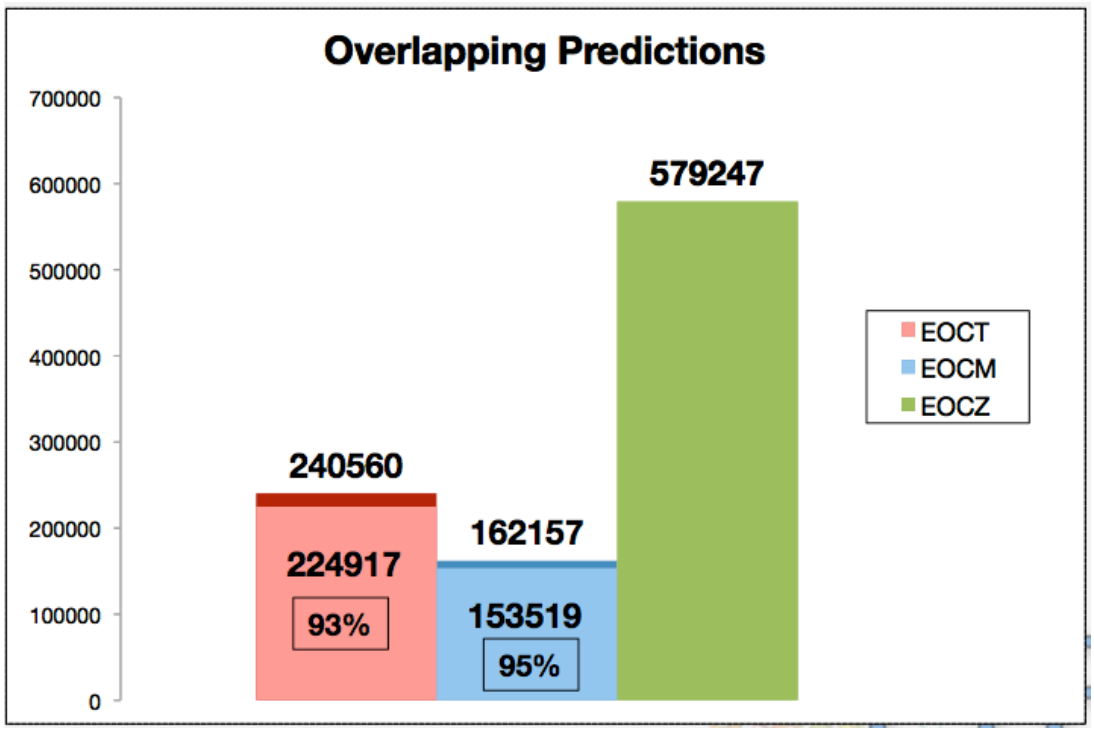


**Figure 2.5: Barchart showing of the overlap of numbers of pairs with predicted scores above 1.0 for the EOCT and EOCM predictors with the EOCZ predictor.** Total numbers of protein pairs with final PIPs scores above 1.0 and the overlap between these predictions for the EOCT (red), EOCM (blue) and EOCZ (green) predictors are shown as an overlapping barchart. Each vertical bar shows the total number of predictions for the predictor it is labelled by, with the number of pairs also predicted by the EOCZ predictor shown in light red (EOCT) and light blue (EOCM), with the percentage of the overlap in the box within each bar. Of the protein pairs in the prediction set, 126,107 were predicted to interact by each of the three methods.

While the cut-off for prediction as interaction versus non-interaction has been designated as a likelihood ratio of 1000.0, the patterns of final scores for the three predictors can be further assessed by comparing numbers of predicted interactions at other cut-off thresholds. Figure 2.6, below, plots these values for the three methods at six threshold between 1.0 and 10000.0.

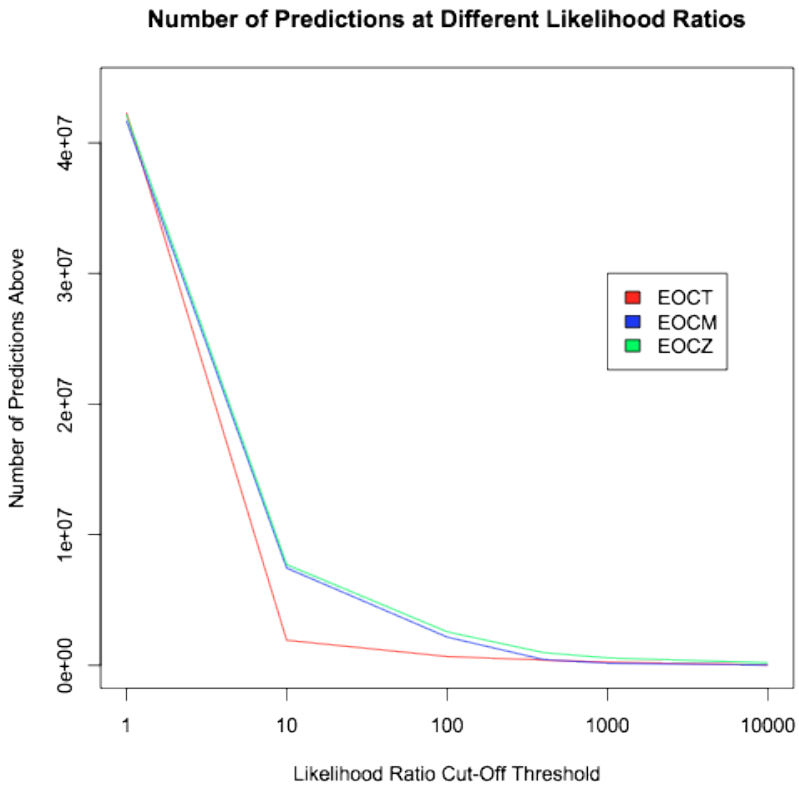


**Figure 2.6: Number of interactions predicted as different likelihood ratio cut-off thresholds.** Total numbers of pairs with final likelihood ratio scores for the EOCT (red), EOCM (blue) and EOCZ (green) are plotted.

While the numbers of predictions by each of the methods above the initial, lower thresholds of 1.0 are similar, as the cut-off values increase, the number of the predictions varies drastically between the EOCT and the EOCM and EOCZ methods. As the final likelihood ratio is calculated as the product of the likelihood ratios of the four individual modules included in the method, this difference could be due to variations in the likelihood ratio values associated with the individual bins in each of the Transitive, Cluster and TransMCL modules. Protein pairs assigned to the highest bin in the Transitive module or to the highest bin in the Cluster module should be assigned the highest bin in the TransMCL module. However, while the likelihood ratios for bins 5 and 6 of Transitive and Cluster modules (212.2 and 175.6, respectively) are similar, the

likelihood ratio for bin 5-5 of the TransMCL module (9946.3) is almost nine times the magnitude. As a result, the final total likelihood ratio for the EOCZ method will be considerably higher than the total likelihood ratio for the EOCT or EOCM methods. As the TransMCL module is meant to be a combinatorial method, this follows that a pair predicted by both components should score higher than a pair predicted by only one.

### 2.3.5 Top Scoring Interactions

In order to examine the top scoring interactions for each of the EOCT, EOCM and EOCZ predictors, the 50 protein pairs with the highest final likelihood ratios for each method were selected for comparison. Across the three predictors, this set of 50 pairs was largely the same, with 27 of the pairs identical across the three predictors. Comparing two of the predictors at once, the EOCT and EOCM methods shared 39 predictions, the EOCT and EOCZ methods shared 28 predictions, and the EOCM and EOCZ methods shared 31 predictions. To look further, the top ten highest scoring pairs across all three methods, shown in Table 2.6, below, were analysed more closely.

First, and most noticeably, the interactions predicted in this set contained pairs of what appear to be subunit interactions of the same complex. For example, in the highest scoring pair, PSMA2 and PSMB4 are both proteasome subunits of the alpha and beta type, respectively. Similarly, in the second and fourth pairs, SNRPE, SNRPF and SNRD2 are all small nuclear ribonucleoproteins that potentially are involved in the same molecular process in the same subcellular localisation. Examining the breakdown of module contributions for each of these proteins reveals that the SNRP protein predictions are built upon a moderately strong Expression module likelihood, a very

high Orthology module likelihood and a moderately high Combined module likelihood before the Transitive, Cluster and TransMCL modules are even considered.

| Protein1 | Protein2 | EOCT Score | EOCM Score | EOCZ Score |
|---|---|---|---|---|
| PSMA2 | PSMB4 | 553420 | 457777 | 25934900 |
| SNRPE | SNRPF | 482770 | 231131 | 22624100 |
| JUN | JUNB | 363670 | 300262 | 445880 |
| SNRPD2 | SNRPF | 280511 | 134297 | 13145600 |
| CGL1 | ACTG1 | 241254 | 167631 | 1441120 |
| CDC2 | CDC25C | 214595 | 199198 | 1080840 |
| PSMB3 | PSMA2 | 199360 | 164907 | 9342540 |
| PSMA2 | PSMA4 | 199360 | 164907 | 9342540 |
| STAT5A | STAT5B | 141998 | 22970.3 | 637127 |
| JUN | JUND | 121959 | 61631.5 | 2740030 |

**Table 2.6: Ten highest scoring interactions for the EOCT, EOCM and EOCZ PIPs predictors.** The EOCT, EOCM and EOCZ score is given for each interaction after division by 1000.0 to adjust for the prior odds ratio.

Overall, this congruence of highest scoring predictions suggests that all three methods are capable of predicting similar sets of plausible interacting pairs.

### 2.3.6 Comparison of PIPs v. 3.0 to PIPs v. 1.0 and 2.0

As the changes in PIPs v. 3.0, excluding the addition of the TransMCL module, did not alter the algorithm substantially and were intended primarily as data updates, the performance of this new version of PIPs should not vary drastically from the previous PIPs v. 1.0 and 2.0. To ensure that the integrity of the predictor has remained, the performances of the EOCT, EOCM and EOCZ predictors on the blind test set were

compared to the EOCT and EOCM predictors in PIPs. Unfortunately, as protein pairs were reassigned into new datasets for the updated predictor, it cannot be guaranteed that this is a true blind test for the previous versions. Therefore, it is possible that results for the PIPs v. 1.0 and 2.0 predictors may be biased.

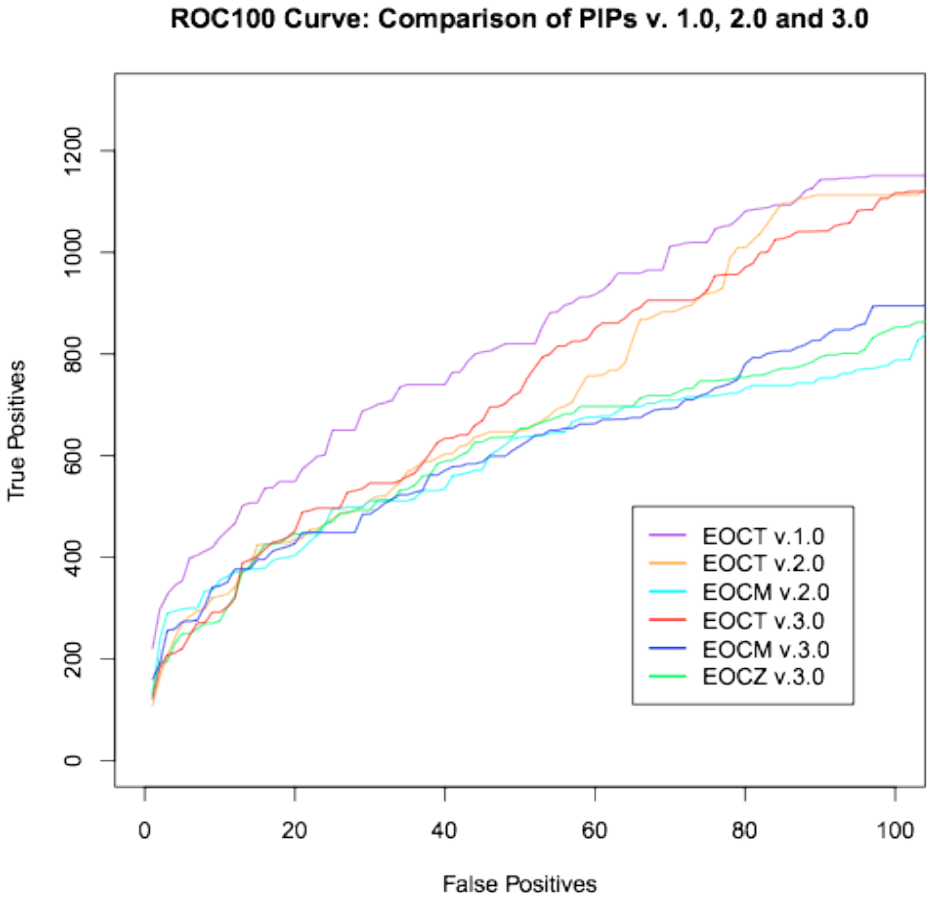


**Figure 2.7: ROC Plot comparing performance of PIPs v. 2.0 to the updated PIPs v. 3.0.** The number of true positive results attained before the first 100 false positive results from a blind test with 2588 positives and 2588 negatives is compared for PIPs v. 1.0 (EOCT, purple), v. 2.0 (EOCT, orange and EOCM, cyan) and v. 3.0 (EOCT, red, EOCM, blue and EOCZ, green).

Figure 2.7, above, shows the ROC100 curves for the five predictors. While the EOCT method in PIPs v. 1.0 (purple, 1151) predicts a consistently higher number of true

positives before the 100th highest scoring false positives than the EOCT methods of v. 2.0 (orange, 1113) and v. 3.0 (red, 1117), the difference is only 38 and 34 predictions. As expected, the EOCM methods in v. 2.0 (cyan, 788) and 3.0 (blue, 895) and the EOCZ method in v. 3.0 (green, 853) all perform worse than their EOCT counterparts. The similarity of these values suggests that PIPs v. 3.0 performs as comparably, as expected, to v. 1.0 and v. 2.0.

### 2.3.7 Performance on Prediction of Known Interactions

In order to assess how well PIPs is able to identify known interactions, prediction scores were obtained for the pairs of proteins in the I2D (OPHID), DIP, HPRD and IntAct databases for each of the EOCT, EOCM and EOCZ methods (Table 2.7, below).

| Database | EOCT | EOCM | EOCZ | Total Interactions |
|---|---|---|---|---|
| **I2D (OPHID)** | 5027 (6.2%) | 3931 (4.8%) | 8030 (9.9%) | 81262 |
| **DIP** | 140 (11.5%) | 148 (12.2%) | 216 (17.8%) | 1215 |
| **HPRD** | 2768 (7.5%) | 2297 (6.2%) | 4339 (11.8%) | 36861 |
| **IntAct** | 669 (3.9%) | 490 (2.8%) | 1020 (5.8%) | 17374 |

**Table 2.7: Number of known protein pairs predicted by PIPs.** Exact numbers of predictions from the PIPs EOCT, EOCM and EOCZ methods along with percentages of protein pairs included in the I2D, DIP, HPRD and IntAct databases with final scores above 1000.0 (prior to adjustment for the prior odds ratio). Self-interactions have not been included.

Of the three PIPs methods, the EOCZ method has predicted a greater number of interactions that overlap each of the four databases considered. While the percentages

of identified interactions are similar to what has been seen with the two previous versions of PIPs, the increased percentages for the EOCZ predictor do suggest that the combination of the Transitive and Cluster modules, though less accurate overall, has increased the coverage of prediction of known positive interactions (Scott & Barton, 2007; McDowall, 2011).

## 2.4 Discussion

After updating the data within the Interactome database to enable PIPs to remain current with the field, the initial aim of further development was to improve the efficiency of the predictor by consolidating the Transitive and Cluster modules into one, eliminating the need for two separate versions of the predictor to attain a full set of predictions. However, while the TransMCL module and resulting EOCZ predictor did increase the total number of predictions to include the majority of those predicted by the EOCT and EOCM predictors on their own, the EOCZ predictor performed slightly worse than the EOCT predictor on multiple blind tests.

While the total number of EOCT predictions is lower than the number of EOCZ predictions, which is the greater compromise - predicting more interactions with less accuracy or predicting less but more accurate interactions at the sacrifice of missing potential positives - must be considered. As the ultimate goal of designing a protein interaction predictor is to ease the experimental process required for validation of interactions and provide a reliable alternative to lab-based methods, having an accurate predictor is more important than one that identifies a large numbers of slightly possible but not highly probable protein pairings.

When analysing the Transitive and Cluster components of the TransMCL module individually, it became apparent that each was performing as it did on its own; however, despite the strong capability of the Transitive portion to score and bin positive and negative interactions appropriately, the Cluster portion appeared to have a detrimental effect. As a result, the module will be kept as an option in PIPs as it does still identify

unique interactions. The consistently stronger performance of the EOCT module, both in previous and the new, updated version of PIPs, above the EOCM module suggests that the simplicity of the transitive scoring method is more effective in assessing potential protein interactions. Therefore, the EOCT method will remain the primary PIPs predictor, with the EOCZ and EOCM methods offered as additional running options.

## 2.5 Conclusions

1) A set of 299 new proteins added over the past two years to the UniProtKB reviewed human protein dataset have been added to the PIPs database along with any available data relevant to the PIPs module.

2) The protein identifiers in the PIPs database have been updated from their now-deprecated IPI identifiers to one of seven alternate identifier reflecting the original source for the protein's inclusion in the IPI database. New cross-references to each of the proteins have been added to the PIPs database.

3) The PIPs Interactome database has now been updated to include the more recent data for orthologues, GO term annotations, post-translational modifications and PFAM domains.

4) The Cluster module has been slightly updated to process data more efficiently, resulting in a dramatic performance and runtime increase.

5) The TransMCL module has been developed through a full Bayesian approach to merge the Transitive and Cluster modules into one module. While this was hoped to increase the performance of the predictor and number of predicted interactions, the new EOCZ predictor still performed worse than the EOCT predictor. While the predictor

will remain as an optional method for prediction, it will not be incorporated as the final PIPs predictor.

# Chapter 3

## PIP'NN: A Neural Network Predictor of Protein Interactions

### Preface

This chapter details the development of an artificial neural network for protein interaction prediction that builds upon the data incorporated in the Bayesian PIPs predictor. The process of dataset, learning method and network structure selection and training the predictor is described in detail.

# 3.1 Introduction

Until this point, all previous development of the PIPs framework has relied on naïve Bayesian statistics. While the clear-cut stages of the Bayesian network strategy offer many advantages, the most notable is the superficial nature of the likelihood ratio calculation that allows results to include not only the final prediction output, but also the details of the contribution of each source of evidence to that prediction. However, the ultimate goal of developing a protein interaction predictor is to produce a tool that performs consistently with the highest possible accuracy. As several other types of machine learning prediction methods exist that would support a protein interaction predictor, it is possible that relying on this Bayesian approach, without investigating these options, is limiting the successfulness of PIPs' predictive capability.

Currently, most predictors of protein-protein interactions rely primarily on either a Bayesian framework or a Support Vector Machine (SVM) methodology and have shown varying levels of success with the two methods (see Chapter 1.4 and 1.5). However, a third learning method, the artificial neural network, has not yet been implemented in the large-scale prediction of human protein-protein interactions. Across the computational biology field, various neural network learning methods have been incorporated to varying degrees in the prediction of the subcellular protein localisation (Emanuelsson *et al.*, 2000; Bodén & Hawkins, 2005; Mooney, Wang & Pollastri, 2011), specific protein interaction sites (Fariselli *et al.*, 2002; Ofran & Rost, 2003; Hamilton *et al.*, 2004), nucleolar localisation signals (Scott, Troshin & Barton, 2011), specific domain recognition (Knisley & Knisley, 2011) and protein structure prediction (Cole, Barber & Barton, 2008) among others.

Naïve Bayesian networks and neural networks differ in how they process data. Most crucially, the fundamental assumption underlying naïve Bayesian networks is that the each source of data, or evidence, is independent. As a result, each piece of evidence is considered separately to calculate the likelihood, in the context of PIPs, that an interaction would occur based only on that evidence, until the calculation of the final likelihood ratio, where these likelihoods are combined. Neural networks, on the other hand, take as input the set of different pieces of evidence and classify the data, from the start, as a whole, rather than as individual entities. By analysing sets of evidence for patterns, neural networks are better able to classify noisy data. With the scarcity of data on the human interactome and lack of available evidence for both individual and pairs of proteins, this strategy of pattern recognition lends itself well to protein-protein interaction prediction.

# 3.2 Methods

## 3.2.1 Data Collection

### 3.2.1.1 Raw Scores Method

For consistency, the raw data considered in the Bayesian version of PIPs (referred to as 'Bayesian PIPs' throughout this chapter) was also used to generate pattern files for training and testing the SNNS neural network. Data values for the Expression and Combined modules supplied to Bayesian PIPs before training, testing and predicting were divided into five values - two Expression scores (the Pearson's Correlation Coefficient and the Spearman's Correlation Coefficient, not assessed in the current version Bayesian PIPs in favour of the Pearson's measure but available) and three Combined scores (the Domain score, PTM score and GO score, all as calculated by assessing the characteristics of these features in the set of known positive and negative interactions).

While these five scores are straightforward, linear calculations representing the relationship between the two proteins in the pair (i.e. correlation for the Expression scores, a calculated Chi-squared score for the Domain score and similarity scores for the PTM and GO scores), dealing with scoring the Orthology module proved more difficult. In the Orthology module, orthologues for the two proteins in the query pair in human, yeast, worm and fly are determined by InParanoid scores, and then these orthologues are assessed for any known interactions within their species. Pairs are then grouped according to two criteria: 1) if one, both or neither proteins have orthologues and 2) if these orthologues interact. Since the binning and scoring in the Orthology module is done discretely rather than by a calculated score, it was decided to take the likelihood

ratio assigned to each bin as the Orthology score to incorporate into the neural network input. While this value is not an absolute measure of the strength of the orthologous relationship, the likelihood ratio values for the module covered a range of values corresponding to their associated bins; for example, bins with only one orthologue and no interaction scored low (0.85) while bins with two orthologues and interactions observed in two or more other species scored much higher (1534.0), with bins with two orthologues and an interaction in one species scored in between (85.19 - 132.16). As such, pairs with interacting orthologues are scored high, while pairs with no interacting orthologues are scored much lower.

In total, considering these sources of evidence resulted in six input data values representing the Expression, Orthology and Combined modules in SNNS PIPs.

#### 3.2.1.2 Likelihood Ratios Method

As a comparison, a second method of input score presentation was also tried. Rather than taking the raw scores analysed by each module in Bayesian PIPs, the final likelihood ratios for each of the modules (Expression, Orthology, Combined, Transitive, Cluster and TransMCL) from the most recent, retrained version of PIPs were taken as input values for the neural network, resulting in a total of six input nodes.

### 3.2.2 Data Normalisation

In order to maintain scoring consistency between the different sources of data, each of the six raw scores was normalised to a value between 0.0 and 1.0 with the standardisation method. The Spearman's and Pearson's Correlation Coefficients, which

range from -1.0 to 1.0, were increased by +2.0 and then divided by 2.0, the maximum score in the set. The Orthology, Domain, PTM and GO term scores were each adjusted by dividing the value by the maximum value in the set (see Equation 3.1, below). Similarly, each of the six likelihood ratios were normalised to values between 0.0 and 1.0 by dividing each score by the maximum likelihood ratio for that source of evidence.

$$norm_i = \frac{score_i}{score_{\max}}$$

**Equation 3.1: Standardisation equation.** $Norm_i$ is the final normalised score, *score* is the original score and $score_{max}$ is the maximum score in the entire set of scores.

### 3.2.3 Datasets

Neural networks typically train more efficiently than Bayesian networks; therefore, the sizes of the training and testing datasets can be smaller than those used in training Bayesian PIPs. Several different training set compositions were tried to compare the impact that the size of the dataset and composition of the dataset would have on how well the neural network trained and predicted.

To maintain consistency with Bayesian PIPs, the neural network version of PIPs was trained and tested with five-fold cross-validation on datasets comprised of a random sample of pairs from the existing positive and negative datasets, giving six data subsets with sets one through five for cross-validation and set six as a blind test. Then, to generate the full network required for generating predictions, the neural network with each specific learning method was trained on datasets one through five altogether.

Therefore, full training set sizes included five times the number of positives and five times the number of negatives per round.

To determine the most effective dataset size and composition for training the networks, three different combinations of positive and negative dataset sizes were tried and are summarised in Table 3.1 and below.

| Pattern File | Positives per Round | Negatives per Round | Normalisation Method | Filtering Details |
|---|---|---|---|---|
| EqualLarge | 5000 per round (25,000 total) | 5000 per round (25,000 total) | Standardisation | No filtering |
| EqualFiltered | 1000 per round (5000 total) | 1000 per round (5000 total) | Standardisation | 3-6 input values > 0.0 |
| EqualFam | 500-600 per round (3253 total) | 1000 per round (5000 total) | Standardisation | 1+ input values > 0.0<br>Split training and testing pairs by superfamilies |

**Table 3.1: Details of datasets tried in training SNNS PIPs.** Three different data subsets were initially tested for training SNNS PIPs. Details of the sets are provided in the table below. Each dataset was derived by taking a random sampling of the six data subsets from Bayesian PIPs, such that sets one through five were implemented during five-fold cross-validation and in training the full, final network and set six was a blind test set. The numbers of positives and negatives in each subset are shown per round with the total number of pairs for the final training shown in parentheses. Each value in the datasets was normalised through standardisation to be between 0.0 and 1.0.

Each of the three datasets was initially constructed by randomly selecting a subset of pairs from the original PIPs datasets (see Chapter 2.2.2: Positive and Negative Dataset Reconstruction). For the EqualLarge dataset, 5000 positives and 5000 negatives were selected for each round of cross-validation. The chosen pairs were kept unfiltered. For the EqualFiltered dataset, after selecting an initial random, larger set of pairs, the dataset

was filtered to only include 1000 positive and 1000 negative pairs per round with between three and six input values above 0.0.

The EqualFam dataset was designed as a more rigid test to ensure the neural network was not over-learning how to classify protein pairs from certain families. To create these training and testing datasets, data on superfamily groupings for individual proteins was first downloaded from UniProtKB/Swiss-Prot and mapped to 15,351 proteins in the PIPs database. This set of 4739 distinct superfamilies was then randomly divided into equal 'training' and 'testing' groups. To assemble the EqualFam datasets, the datasets from the original PIPs were filtered in two steps: first, only those pairs that had both proteins with assigned superfamilies in the 'training' set were selected, and then the resulting subset was filtered to contain only those pairs with between three and six scores above 0.0. While selecting negative examples for this dataset was straightforward, selecting positive pairs was limited first by the smaller number of pairs to consider and then by the possible 'train-train' and 'test-test' superfamily pairing. As a result, despite lowering the minimum number of input scores for filtering to one, each of the cross-validation subsets was still less than 1000, and the final training set contained 3253 positive and 5000 negative pairs.

### 3.2.4 SNNS

SNNS (the Stuttgart Neural Network Simulator) is a package developed in the late 1990s that incorporates over 20 different neural network learning methods with multiple parameter options to train and test combinations of learning methods, training datasets and network structures with ease (Zell, 1995). In addition to offering this wide range of

options for network constructing and training, SNNS includes a script that packages the trained network into an executable ANSI C program that can compute output values for new input patterns as a standalone program without the rest of the SNNS framework.

The individual components of SNNS implemented are described in more detail in the sections following.

## 3.2.5 SNNS Set-Up

### 3.2.5.1 Pattern Files

SNNS requires a specific format for data presentation to the neural network in the form of a 'pattern file' containing a header with information about the data being presented and the input and output structure of the neural network followed by a body section, where the information for each pair is presented as a set of numerical values (a pattern) followed by a numerical value for its expected output (see Figure 3.1 for an example). In order to both train and test the neural network, two pattern files were written for each round of cross-validation: one with the training set (four of the datasets) and one with the testing set (one of the datasets), thus requiring a total of ten pattern files for complete validation. To train and test the full, final network, two additional pattern files were written: one with sets one through five for training and one with set six as a blind test.

Pattern Header

Protein Pair
Input Scores
Expected Output

```
SNNS pattern definition file V3.2
generated at Fri Aug 10 15:38:09 2012

No. of patterns : 2000
No. of input units : 6
No. of output units : 1

# 2-680
0.56729 0.611885 0.0 0.00359177 1.33883E-6 0.00431854
#Output
0
# 3-5638
0.718422 0.716317 0.0 0.0 4.3425E-6 0.0074613
#Output
1
# 3-11749
0.750754 0.727867 0.0 2.03642E-6 8.13086E-6 0.0
#Output
1
```

**Figure 3.1: Annotated example of an SNNS pattern file.** The first portion of the pattern file set-up required for SNNS is shown. All pattern files must contain a header with the number of patterns and the input and output structure of the network that is followed by sets of patterns. Lines with #'s are not read so have been used to annotate each pattern input with the pair it corresponds to and separate the output. Input patterns are given in floating numbers (with or without scientific notation) and have been normalised to values between 0.0 and 1.0.

#### 3.2.5.2 Batch Files

The 'recipe' for training and testing the neural network is presented to SNNS with the 'batch file', which contains information on how the network should be initialised, the order in which patterns should be examined, the neural network method that should be employed and its required parameters, the number of training cycles and the type of output files that should be written (see Figure 3.2 for an example).

*Initalisation Function* - The Initialisation Function sets the weights between connections before training to avoid bias and was set to Randomize_Weights to assign weights between -0.1 and 0.1 to all connections at random.

*Update Function* - The Update Function determines the order in which the connections between input, hidden and output nodes should be revalued. The Topological_Order function, which updates input, hidden and output nodes in order, was chosen to update connection weights.

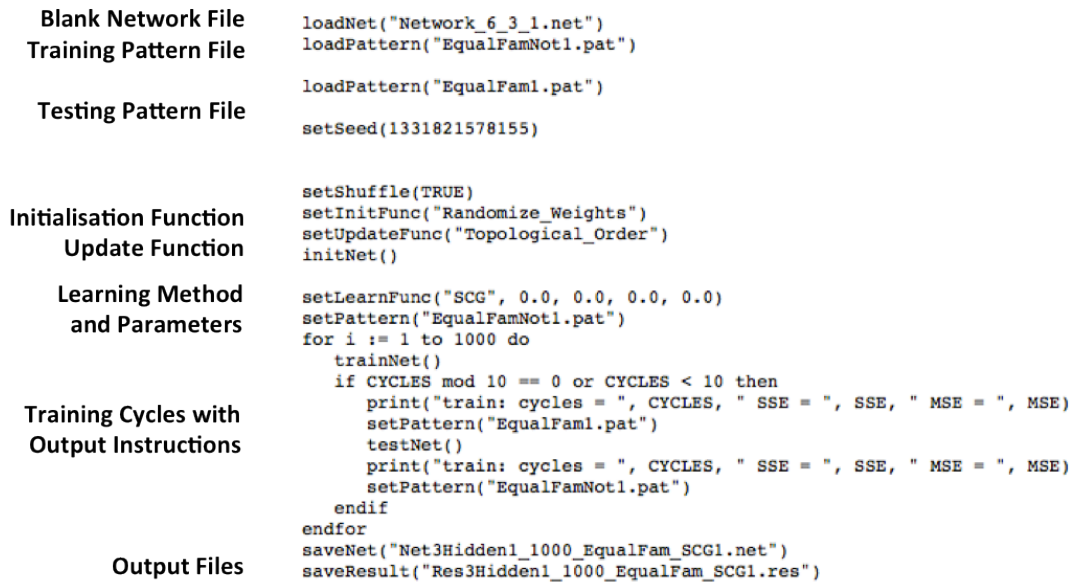


**Figure 3.2: Annotated example of an SNNS Batch files.** An annotated example of a batch file input to the SNNS program for a network with six input, three hidden and one output nodes for cross-validation training on the EqualFam dataset is shown above. Both training (EqualFamNot1.pat) and testing (EqualFam1.pat) patterns are loaded, the network and learning method are initialised, and then the program loops through the set number of cycles (in this example, 1000), calling the trainNet() function. This batch script calls for output to be written to a file given as a commandline argument after each of the first ten cycles and then after every tenth cycle. The weighted network from training is saved (saveNet) along with a file with the results for each prediction in the training set (saveResult).

*Learning Function* - The Learning Function parameter specifies which method of learning should be employed for training. To compare the effects of different learning

methods on the effectiveness of training the neural network, a range of methods were attempted at first until ultimately narrowing down to three for final comparison: Std_Backpropagation, BackpropChunk and SCG (Scaled Conjugate Gradient). The Std_Backpropagation method learns through back propagation (for more detail, see Chapter 1.5.4.3.3: Back propagation) with errors calculated and weights reassigned after all patterns are processed during each training cycle. The BackpropChunk method is a variation on back propagation in which connection weights are revalued after presenting a chosen number (or a 'chunk') of patterns and evaluating the error, allowing the network to be updated more frequently according to a range of inputs. The Scaled Conjugate Gradient method is an additional feed forward learning method that, rather than proceeding through the network as a gradient and minimising the error step-by-step, goes through to adjust weights based on both what their values are before and after analysing the pattern (for more detail, see Chapter 1.5.4.3.4: Scaled Conjugate Gradient).

*Cycles/Epochs* - Different amounts of training cycles (epochs), ranging from 250 to 1000, were tried to determine how quickly the network could be trained and assess if any overtraining was occurring at any point. Ultimately, 1000 cycles was chosen for training in all learning methods.

#### 3.2.5.3 Network Files

In order to set the base structure of the neural network, SNNS requires a blank 'network file' that presents the number of input, hidden and output nodes, draws the map of their connections and sets all connection weights to zero. After training, this network file is rewritten with new weight values for the connections and is compiled into an executable

ANSI C program that can predict individual outcomes through the neural network (see Figure 3.3 for comparative examples of the un-weighted and weighted networks).

A three-layer network was implemented for each of the learning methods chosen. The input layer in each network contained six nodes with each neuron corresponding to the raw score from an individual source of evidence or the likelihood ratio for each module (for further detail on data selection, see Chapter 3.2.3: Dataset Collection, above).

The output layer for each network consisted of one node corresponding to a final output score between 0.0 and 1.0, such that values above or below a set threshold would indicate either a positive (Predicted Interaction) or negative (No Predicted Interaction) prediction, respectively.

While there have been previous suggestions about the optimal number of hidden nodes for a network (described in Basheer & Hashmeer, 2000), there is no one universally perfect value; therefore, a range of hidden node values between 1 and 100 were tested for each of the learning functions. After several rounds of analysis of the effects of different numbers of hidden nodes on training effectiveness between networks and learning functions, five representative numbers of hidden nodes (3, 12, 50 and 100) were chosen for in depth comparison.

A.

```
SNNS network definition file V1.4-3D
generated at Thu Jul 12 11:39:09 2012

network name : PIP_net
source files :
no. of units : 57
no. of connections : 350
no. of unit types : 0
no. of site types : 0


learning function : Std_Backpropagation
update function   : Topological_Order


unit default section :

act      | bias     | st | subnet | layer | act func      | out func
---------|----------|----|--------|-------|---------------|-------------
0.00000  |  0.00000 | h  |      0 |     1 | Act_Logistic  | Out_Identity
---------|----------|----|--------|-------|---------------|-------------


unit definition section :

no. | typeName | unitName | act      | bias     | st | position | act func | out func | sites
----|----------|----------|----------|----------|----|----------|----------|----------|-------
  1 |          | unit     |  0.00000 |  0.00000 | i  |  1, 1, 0 |||
  2 |          | unit     |  0.00000 |  0.00000 | i  |  1, 2, 0 |||
  3 |          | unit     |  0.00000 |  0.00000 | i  |  1, 3, 0 |||
  4 |          | unit     |  0.00000 |  0.00000 | i  |  1, 4, 0 |||
  5 |          | unit     |  0.00000 |  0.00000 | i  |  1, 5, 0 |||
  6 |          | unit     |  0.00000 |  0.00000 | i  |  1, 6, 0 |||
  7 |          | unit     |  0.00000 |  0.00000 | h  | 23, 1, 0 |||
```

B.

```
SNNS network definition file V1.4-3D
generated at Fri Aug 10 16:32:04 2012

network name : PIP_net
source files :
no. of units : 57
no. of connections : 350
no. of unit types : 0
no. of site types : 0


learning function : SCG
update function   : Topological_Order


unit default section :

act      | bias     | st | subnet | layer | act func      | out func
---------|----------|----|--------|-------|---------------|-------------
 0.00000 |  0.00000 | h  |      0 |     1 | Act_Logistic  | Out_Identity
---------|----------|----|--------|-------|---------------|-------------


unit definition section :

no. | typeName | unitName | act      | bias     | st | position | act func | out func | sites
----|----------|----------|----------|----------|----|----------|----------|----------|-------
  1 |          | unit     |  0.67431 | -0.63012 | i  |  1, 1, 0 |||
  2 |          | unit     |  0.72911 |  0.30698 | i  |  1, 2, 0 |||
  3 |          | unit     |  0.00000 |  0.68236 | i  |  1, 3, 0 |||
  4 |          | unit     |  0.00223 |  0.92682 | i  |  1, 4, 0 |||
  5 |          | unit     |  0.00001 | -0.95978 | i  |  1, 5, 0 |||
  6 |          | unit     |  0.00741 |  0.69305 | i  |  1, 6, 0 |||
  7 |          | unit     |  0.50780 | -0.08646 | h  | 23, 1, 0 |||
```

**Figure 3.3: Examples of an unweighted and weighted network file.** Comparison of an unweighted network file (A) before training and the resulting weighted network file (B) post-training.

## 3.2.6 Training and Assessing Training Success

### 3.2.6.1 Five-Fold Cross-Validation and Parameter Selection

Five-fold cross-validation training and testing was completed on each combination of pattern file/dataset, learning method and network and was initiated by running the *batchman* batch script in the SNNS training package by providing the batch file containing the detail for training and a name for the output file.

During training and testing, SNNS was set up to return the sum of standard error (SSE), the sum of the differences between the expected outcome (either 0.0 for negatives or 1.0 for positives) and the predicted outcome (any value between 0.0 and 1.0) for each pattern in the training dataset, for the first ten and then every tenth training and testing cycle in the round. These values were then plotted as curves to analyse how quickly the SSE decreased, when it levelled out and if it began to increase at any point during the training. The best combination of parameters, hidden and output nodes for each of the three learning methods with the lowest overall SSE and smoothest curves that displayed clear evidence of training (as indicated by more than one sudden drop in the SSE during the early rounds and no further increase in the SSE over the course of the training) were selected for further analysis and prediction. Following cross-validation, nine combinations of parameters (with one for each of the three learning methods for each of the three datasets) remained for full training and testing in the raw scores and likelihood ratios methods. In order to assess the distribution of scores each neural network assigned during training, histograms for the positive and negative pairs in the training dataset were plotted.

#### 3.2.6.2 Full Training

For full training and testing, the selected combinations were re-trained on datasets one through five and tested on the previously unseen blind test set. To assess the accuracy of this training and if there was a clear divide between output scores for the known positives and negatives, SNNS allows the option of returning a 'results file' with the output scores assigned to each pattern during training. These expected outcomes (either 0.0 for negatives or 1.0 for positives) and the SNNS predicted outcomes (a floating number between 0.0 and 1.0) were then compared for each pair and the predicted outcomes plotted as a histogram to analyse the distribution of scores within the datasets and determine if there was a clear cut-off threshold for positive and negative scores. Additionally, the true positive (TPR) and false positive rates (FPR) and Matthew's Correlation Coefficient (MCC), a measure of the error between two sets, were calculated at a series of thresholds. Finally, the FPR and TPR were plotted, and the result ROC curve analysed for smoothness, rate of increase and area under the curve by the pRoc package for R (Robin *et al.*, 2011).

### 3.2.7 Prediction of Interactions

After training, a new network file was automatically written to include the computed weight values between the input, hidden and output nodes. This network file was then compiled into an executable ANSI C program with the SNNS program *snns2c*, which provided two files, a header file and a C file. A wrapper method was written to read in a pattern for a given pair of proteins, present the pattern to the network and write the final value returned by the neural network into an output file. In order to select a cut-off threshold for prediction, the numbers of interactions predicted at a range of cut-offs

were assessed. Additionally, a plot for Accuracy ((TP+TN)/(Positives+Negatives)) versus Precision (TP/(TP+FP)) was drawn for the EqualFiltered blind test set to identify the cut-off threshold where the values intersected. Ultimately, pairs with final output scores greater than or equal to 0.5 were considered as positive ('Predicted Interaction'), while pairs with scores lower than 0.5 were considered as negative ('No Predicted Interaction').

Finally, output scores were predicted for the full set of protein pairs as in Bayesian PIPs.

### 3.2.8 Incorporation of the Transitive Module from Bayesian PIPs for the Raw Scores Method

In order to mirror the Bayesian PIPs two-stage framework and incorporate the network analysis component into the neural network, a second training step was attempted in the raw scores method following the same methodology as in the Transitive module. After obtaining the predicted output for each protein pair in the PIPs predictive set, two initial, predicted interaction networks were assembled: one with 1,999,770 pairs with scores above a threshold of 0.5 (the '0.5 Network') and one with 517,490 pairs with scores above a threshold of 0.7 (the '0.7 Network'). A transitive score was then calculated for each pair in the predictive set in the same manner as the Transitive module in Bayesian PIPs (see Equation 3.2, below).

$$T = \frac{\sum_{e \in E_c} s_e}{1 + \left| E_i \setminus E_c \right| + \left| E_j \setminus E_c \right|}$$

**Equation 3.2: Transitive Neighbourhood Topology Score.** $E_i$ is the set of edge for protein $i$, $E_j$ is the set of edges for protein $j$, $E_c$ is the set of common edges between proteins $i$ and $j$, and $s_e$ is the likelihood ratio for each common edge between proteins $i$ and $j$.

In order to determine how these predictions could be included in the neural network framework, these new sets of transitive scores were then handled in two ways. In the first method of transitive scoring, under the principle that the neural network first step could determine the initial interaction network, but grouping the interactions in that network further would enhance prediction accuracy, predictions were considered based upon their raw transitive score that was returned from the calculation alone without any further processing.

In the second transitive method, a second, smaller neural network was implemented that considered as input the output score from the first stage of predictions and the raw transitive score normalised to between 0.0 and 1.0. The two networks were trained and tested with six different hidden node values (5, 10, 25, 50, 75 and 100), the SSE curves plotted and the most optimal number of hidden nodes selected as described above. In order maintain consistency with the first stage, the SCG learning method was chosen along with the same set of parameters and initialisation and update functions.

The performance of both methods of scoring at both the 0.5 and 0.7 initial network cut-off thresholds were compared by plotting the full ROC curves for the EqualFiltered blind test dataset with 1000 positives and 1000 negatives not seen during training. Areas under the curve were calculated and compared with Delong's test for two ROC curves using the pROC package for R (Robin *et al*., 2011). Additionally, the distributions of output scores assigned to the positive and negative pairs in the training dataset were plotted as histograms.

# 3.3 Results

Designing and training a neural network requires three main steps:

1) Selection of appropriate training and testing datasets

2) Design of an appropriate network structure that corresponds to the amount of input data presented (input nodes), the desired outcome (output nodes) and how that data should be processed within the network (hidden nodes)

3) Selection of an appropriate learning method and its associated parameters.

Finding the most suitable and successful combination of these three aspects requires multiple stages of testing. Details of the step-by-step process undertaken to develop the final trained neural network are described below.

## 3.3.1 The Raw Scores Method

### 3.3.1.1 Dataset, Learning Method and Hidden Nodes Selection

In order to select the optimal dataset-learning method-hidden node combination for the final PIPs neural network (referred to as 'SNNS PIPs' throughout this chapter), the different combinations of components were tested in a step-by-step manner, starting with three datasets with three learning methods each and four hidden nodes, giving a total of 36 initial networks (referred to from now on as 'combinations'). Figure 3.4, below, shows a graphical summary of the workflow undertaken to arrive at the final SNNS PIPs neural network predictor.

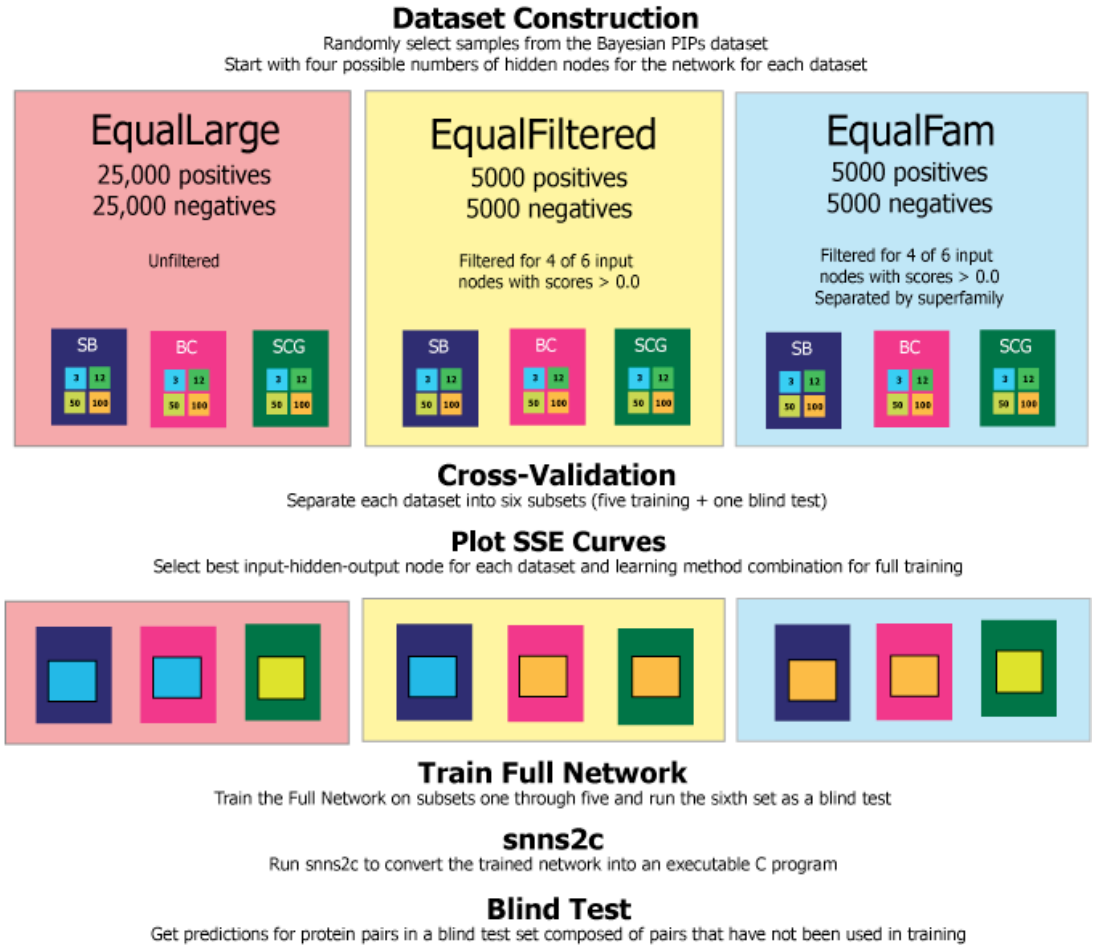


**Figure 3.4: Summary of selection process for the final SNNS dataset, learning method and hidden nodes combination.** The graphical schematic above depicts the general workflow for narrowing down the different combinations of training datasets, learning methods and hidden nodes to the final SNNS PIPs predictor. First, three different training datasets sets were tried (describe in Methods, above) - EqualLarge (pink large box), EqualFiltered (yellow large box) and EqualFam (blue large box). For each training dataset, there were three learning methods, depicted with the mid-sized boxes - Std_Backpropagation (dark blue), BackpropChunk (bright pink) and SCG (dark green). For each learning method, there were four different network structures with varying numbers of hidden nodes, shown with the small, square boxes - three (bright blue), 12 (bright green), 50 (bright yellow) and 100 (orange). During the first stage, each of the 36 potential dataset-learning method-hidden node combinations was trained with cross-validation and each SSE curve plotted. For each dataset-learning method, the network structure with the best SSE curve was chosen, giving nine combinations that were then trained on the full training dataset. After this stage, histograms of the distributions of scores assigned to the training set pairs during training were plotted to determine if that network was capable of distinguishing between positives and negatives. Predictions were then made with each of the remaining combinations for the sixth test data subset as a blind, and ROC curves were plotted to compare the methods. Finally, the best dataset-method combination was selected as the final network for predictions on the full set of possible protein pairs.

#### 3.3.1.2 EqualLarge Dataset

As a first step, it was necessary to assemble the datasets for training and testing each learning method-hidden node combination. The first dataset tried, the 'EqualLarge' dataset, contained a total of 25,000 positives and 25,000 negatives that were selected randomly from the original PIPs datasets and left unfiltered. To determine which network structure was most effective for each of the three learning methods (Std_Backpropagation, BackpropChunk and SCG, for more details see Chapter 3.2.5.2: Batch Files - Learning Methods), networks with four different numbers of hidden nodes (3, 12, 50 or 100) were trained with cross-validation. Next, the average sum of standard error (SSE) for the training datasets across the rounds was plotted for each learning method-hidden node combination, and the number of hidden nodes for each learning method with the lowest overall SSE and smoothest curve were selected for training on the full training set. This selection resulted in one combination for each learning method:

1) Std_Backpropagation with three hidden nodes
2) BackpropChunk with three hidden nodes
3) SCG with 50 hidden nodes.

Next, the output values (ranging between 0.0 and 1.0, where 0.0 would indicate a strong negative and 1.0 would indicate a strong positive) assigned to each pair in the training set were plotted as histograms to assess the distribution of the scores for the known positive and known negative examples. Ideally, successful training of the network would be indicated by two clear distributions, with scores for the negative pairs clustered across the range of 0.0 and 0.5 and the scores for the positive pairs clustered

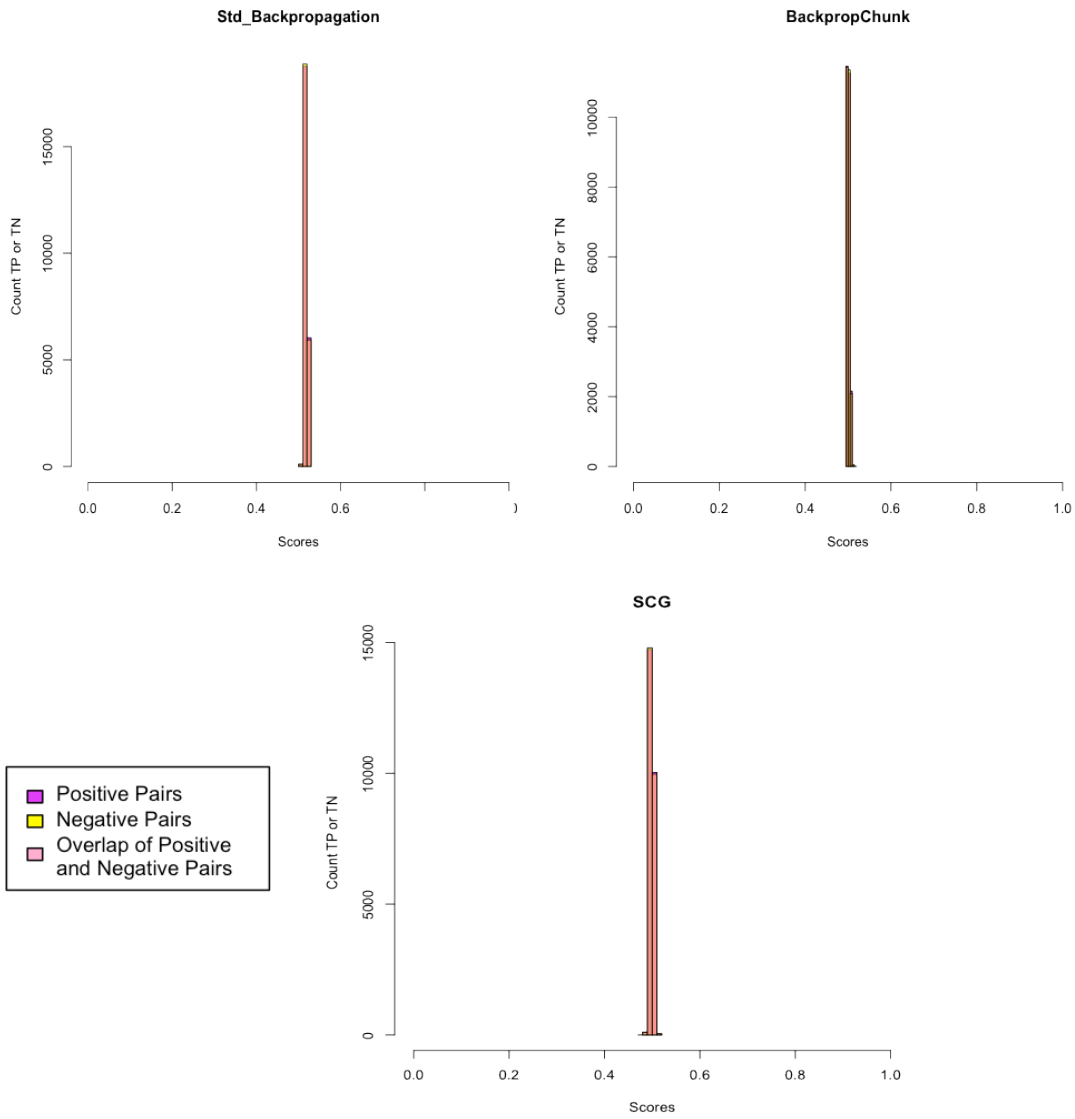


**Figure 3.5: Score Distributions for training the BackpropChunk, Std_Backpropagation and SCG networks with the EqualLarge dataset.** Histogram distributions for the scores during training the BackpropChunk, Std_Backpropagation and SCG neural networks with the EqualLarge dataset, where pink represents the scores for positive pairs and yellow represents the scores for negative pairs. While the output values should range from 0.0 to 1.0, in all three learning method-hidden node combinations, the scores of both the positive and negative datasets are clustered around 0.5, with no significant difference between the distributions of positive and negative scores for any of the methods (KS-test: Std_Backpropagation: p-value = 0.637, D = 0.007, BackpropChunk: p-value = 0.407, D = 0.008, SCG: p-value = 0.880, D = 0.005), suggesting the networks did not train.

across the range of 0.5 and 1.0. Instead, as shown in Figure 3.5, the score distributions for all three learning method-hidden node combinations trained on the EqualLarge

dataset fell into only a narrow range around 0.5 with no distinctive separation between scores in either set.

#### 3.3.1.3 EqualFiltered Dataset

As a result, the EqualLarge dataset was rejected, and a new approach was attempted with the second training set, the EqualFiltered dataset. After examining the EqualLarge dataset, it became apparent that several of the pairs used for training had limited, if any, available data. Speculating that this might be contributing to the network not training, the EqualFiltered dataset was constructed by selecting pairs and filtering them such that every pair had at least three quantifiable values for input, giving a final training network size of 5000 positives and 5000 negatives. In order to determine the individual, optimal network structures for the three learning methods, each of the twelve learning method-hidden node combinations was trained with cross-validation and the average SSE values for the five rounds plotted (Figure 3.6, below).

For each learning method, the network structure with the lowest overall SSE and smoothest curve was chosen for full training, resulting in three learning method-hidden node combinations:

1) Std_Backpropagation with three hidden nodes

2) BackpropChunk with 100 hidden nodes

3) SCG with 12 hidden nodes.

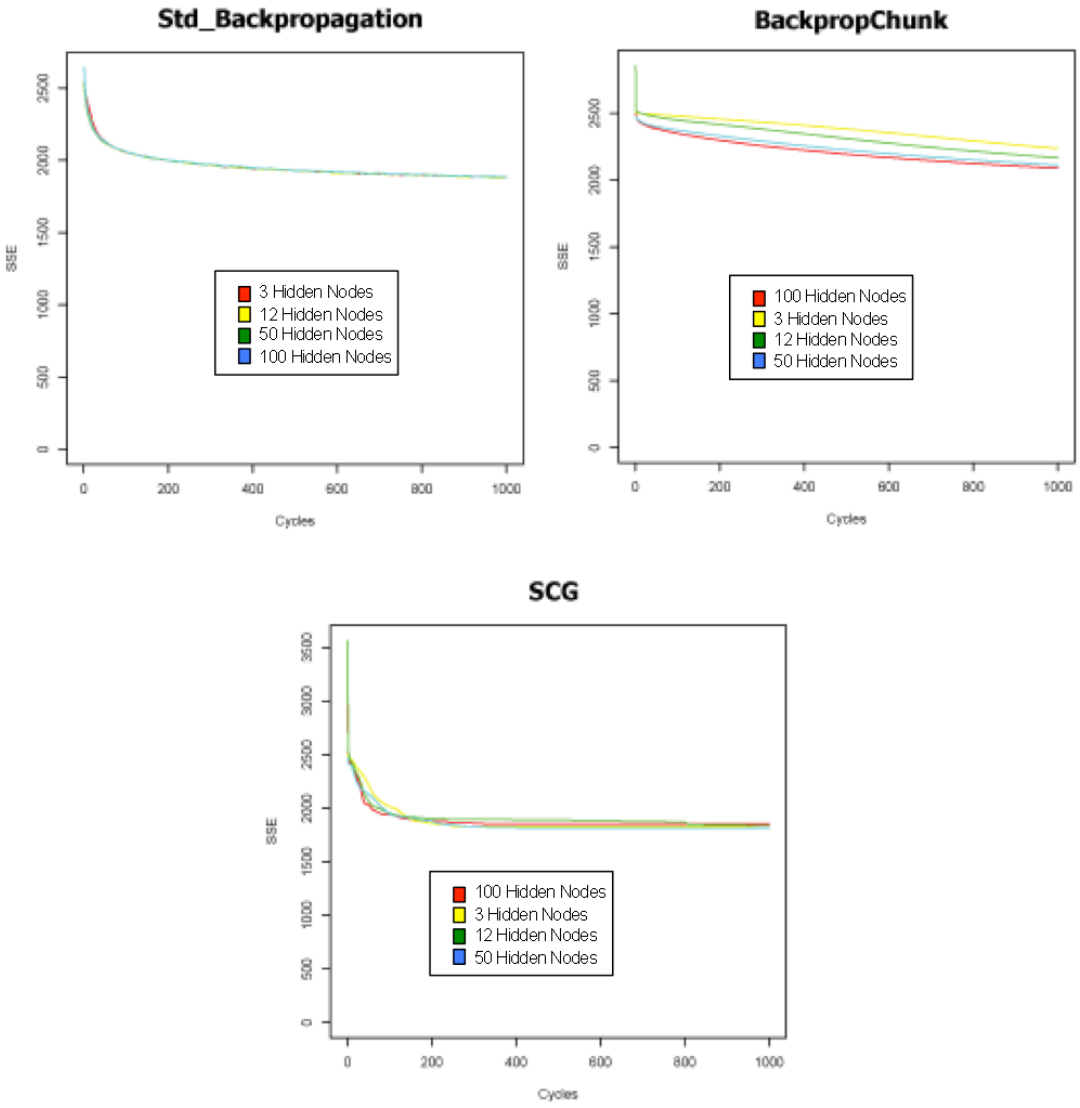


**Figure 3.6: SSE curves for the EqualFiltered dataset.** Plotted SSEs for the four network structures for each of the three learning methods when trained with cross-validation with the EqualFiltered dataset are compared. While the varying numbers of hidden nodes show little difference for the Std_Backpropagation (left) and SCG (right) methods, the network with 100 hidden nodes maintained a lower SSE overall for the BackpropChunk (centre) method. Additionally, the non-smooth profile of the Std_Backpropagation curves, when compared to the curves from the BackpropChunk and SCG methods, suggest that the network is unlearning during the successive cycles and might not be as strong post-training.

The networks for these three combination were then trained on the full set of training data, and the positive and negative score distributions were again plotted as histograms, shown in Figure 3.7, below. Unlike the distributions for the networks in the EqualLarge dataset, each of the networks was able to successfully discriminate pairs in the positive

and negative datasets. Additionally, the range of scores covers the entire possible expanse of 0.0 to 1.0, indicating that network has learned to linearly assign scores according to the strength of the prediction.

#### 3.3.1.4 Blind Test Set

The distributions plotted above are only useful as an initial indication for whether or not the network is capable of assigning realistic scores to the patterns it is being presented with. However, since this analysis is looking only at input vectors that the network has used for learning, it is a biased assessment of how well it can perform when faced with the input values from unseen pairs of proteins. The only way to test how well the network has learned is through running a blind test on a set of pairs not seen during the training. Therefore, predictions were made for a set of 1000 positive and 1000 negative pairs.

The ROC plot in Figure 3.8 compares the performance for the SCG (red), BackpropChunk (green) and Std_Backpropagation (blue) methods. While the curves indicate little difference in prediction accuracy between the SCG and Std_Backpropagation learning methods (AUC of both =0.775), both perform better than the BackpropChunk method (AUC=0.748).

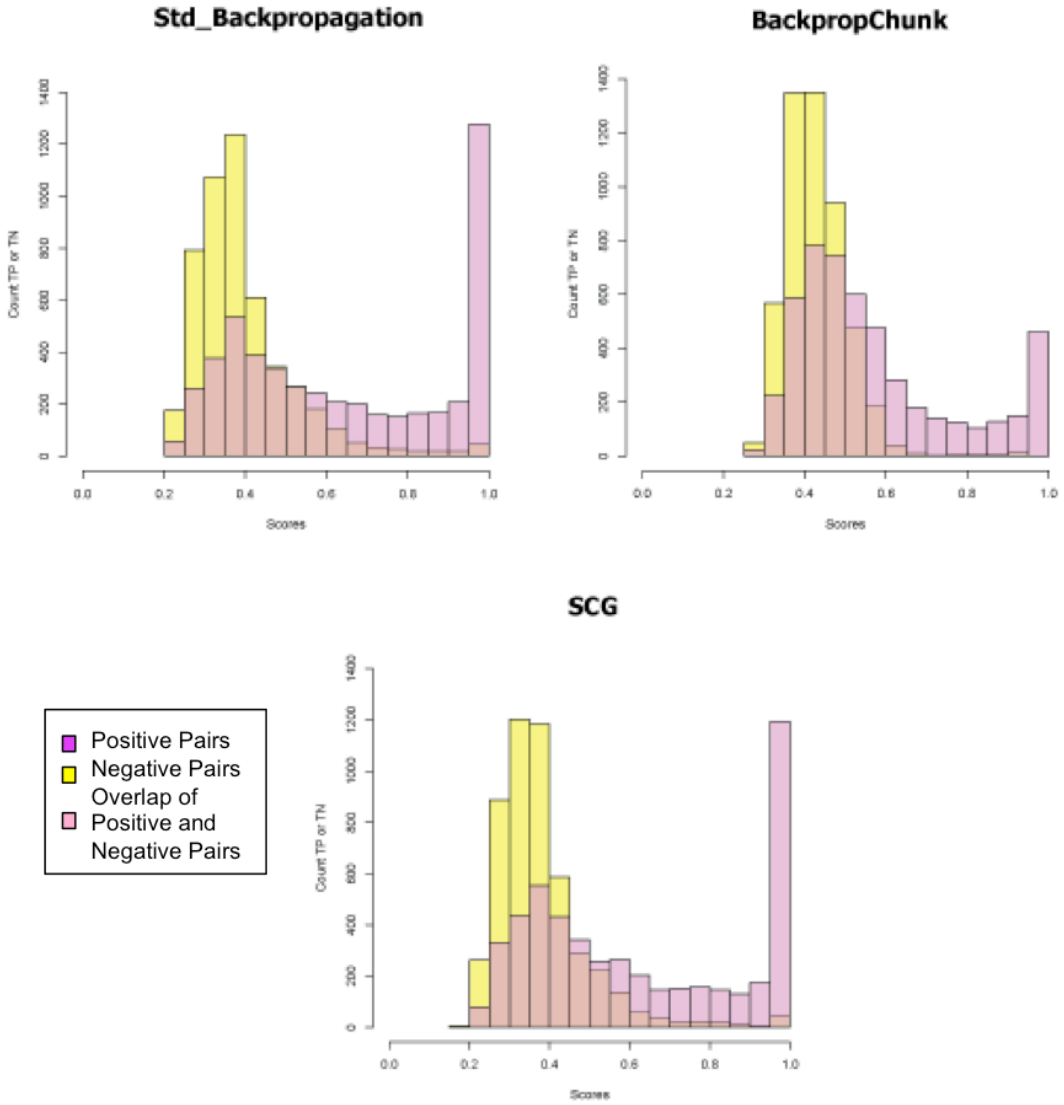


**Figure 3.7: Histogram distributions of scores assigned to the EqualFiltered dataset during training.** The distributions of scores assigned to the positive (pink) and negative (yellow) pairs in the EqualFiltered dataset during training suggest that the Std_Backpropagation (left, p-value = 2.2e-16), BackpropChunk (centre, p-value = 2.2e-16) and SCG (right, p-value = 2.2e-16) methods show a significant difference. This difference in positive and negative distributions suggests that each of the networks has been trained successfully.

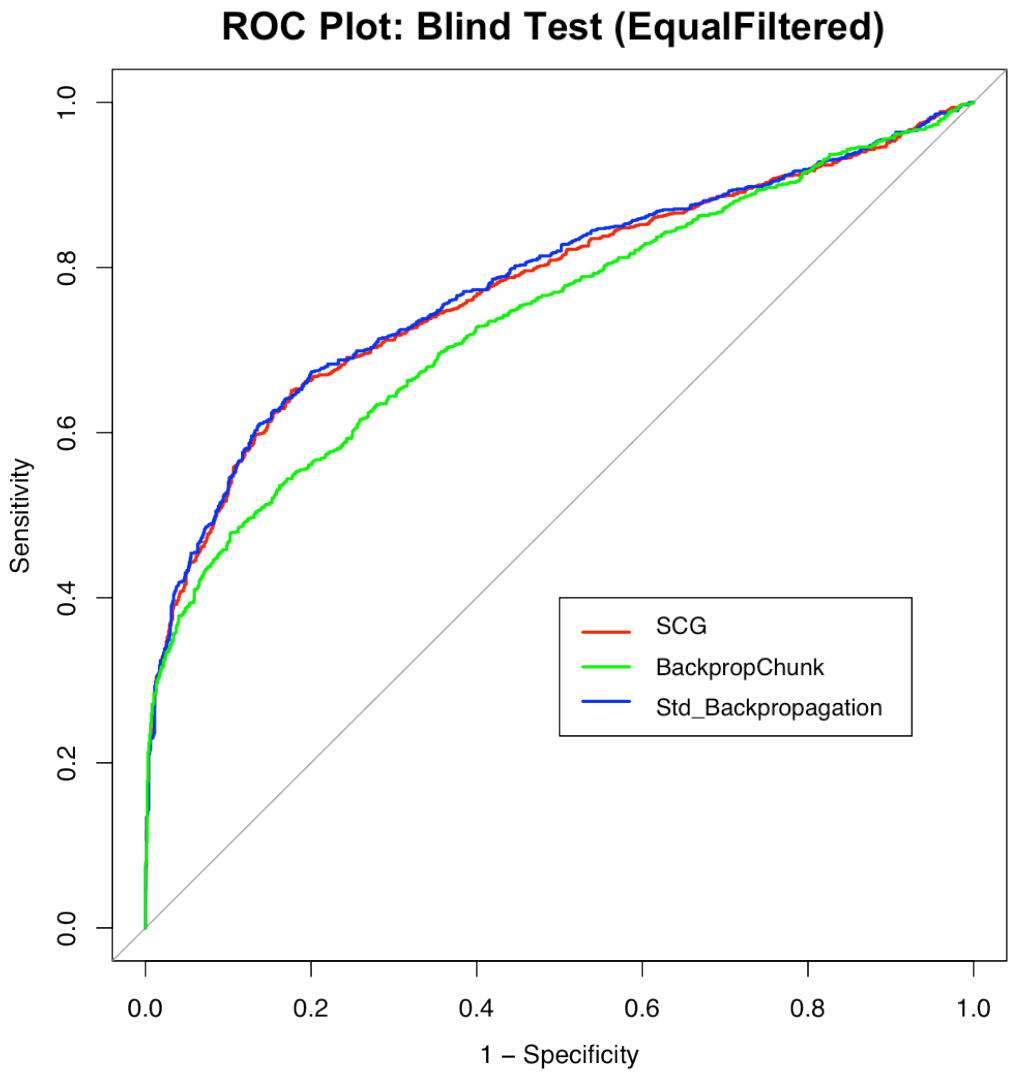


**Figure 3.8: ROC plot for the EqualFiltered blind test set predictions.** ROC plot comparing the prediction accuracy of the neural network predictors trained with the SCG (red, AUC=0.775), Std_Backpropagation (blue, AUC=0.775) and BackpropChunk (green, AUC=0.748) learning methods on a blind test set with 1000 positive and 1000 negative pairs. ROC plots were drawn using the R package pROC (Robin *et al.*, 2011).

Both the EqualFiltered dataset and the blind test set above are comprised of a random sampling of proteins representative of the entire set of potential protein pairs. However, it is possible that the sets, despite their random selection, may still be biased toward

including a subset of proteins that share similar structures or intrinsic properties that have made them more, or less, easily studied and characterised.

In order to ensure that the blind set was not biased, composition of the test set was analysed further. One method of grouping proteins is through classification based on structural properties. There are several sources for this classification, most well-known of which is the SCOP database (Andreeva *et al.*, 2008), which groups proteins into hierarchies with superfamilies as the highest level and proceeding down through families, domains, classes and folds (Conte *et al.*, 2000). Additionally, PFAM groups proteins according to a hierarchy of family, domain, repeat and motif (Punta *et al.*, 2012) and InterPro integrates several different databases to group proteins into superfamily, family and sub-family levels (McDowall & Hunter, 2011; Hunter *et al.*, 2012). While the current Bayesian and SNNS PIPs data includes information from PFAM about domain classifications, it does not incorporate superfamily or family classifications into its evidence.

As superfamilies and families contain subsets of proteins that are similar in structure and therefore, likely similar in physiochemical, sequence or other properties, it is possible that pairs of proteins from within the same superfamily may share similar features. By extension, this similarity may mean that these pairs of proteins share near-identical scores for the sources of evidence incorporated into PIPs. While this similarity does not affect the performances of the PIPs predictors in the context of the entire proteome, it does mean that, even with random selection, the subset could be biased to include a higher or lower number of pairs from certain combinations of superfamilies. For example, proteins in superfamily or family groups that are studied and visualised

more easily due to their intrinsic properties may have not only more evidence available, but also have more known interactions.

As a result, it is possible that the positive blind test set contains a large number of proteins from the same set of classifications about which much is known, while the negative test set contains proteins belonging to less well-characterised superfamilies and families. If this over- and under-representation of certain groups of proteins is the case, the neural network could be being trained and tested on positives interactions comprised only of similar proteins. This overtraining could lead to a predictor that is overly sensitive to only certain positive and negative input patterns and is unable to handle patterns from protein pairs unlike those it has been trained on.

In order to investigate this potential bias, the superfamily composition of the blind test set was analysed. Since PIPs already contains evidence from InterPro, the domain classifications for each protein in the test set were attained and compared against the composition of the domain classifications for the entire set of proteins in PIPs as a whole. When compared, the percentages of the InterPro families in the blind test set versus the percentages of the InterPro families in the whole PIPs protein set showed no significant difference (Pearson's Correlation Coefficient = 0.773, t =58.20, df = 2283, p-value = 2.2 $e^{-16}$) (see Figure 3.9, below). Despite clear outliers, the distribution of families in the blind set when compared to the set of all proteins is moderately correlated, suggesting that the blind set is a representative subset of the entire protein set.

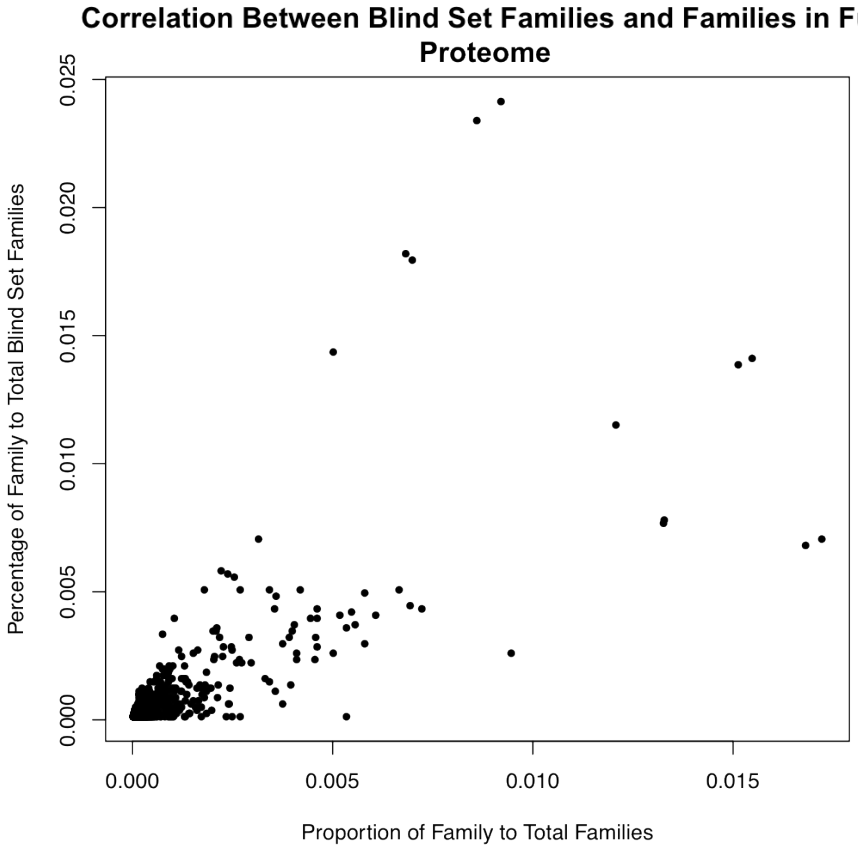


**Figure 3.9: Correlation between proportion of the times families are seen in the EqualFiltered blind test set and proportion of times families are seen across the full proteome.** The percentage a family is seen across the full PIPs protein dataset (x-axis) is compared to the percentage that that family appears in the blind test set (y-axis). Pearson's Correlation Coefficient = 0.773, df = 2283, t = 58.20, p-value = 2.2 e 16) values were calculated through R.

#### 3.3.1.5 EqualFam Dataset

Despite the reassurance that the blind set was a representative sample of the whole set of potential interacting protein pairs, it is still possible that the neural network predictors are over-learning how to predict on proteins with certain properties and then performing well because they are being tested against proteins with those same properties. Therefore, as a further level of testing, a new dataset, EqualFam, was constructed to

contain only pairs where each of the individual proteins belonged to a member of the designated 'training' set of superfamilies.

Following the same selection process as the EqualLarge and EqualFiltered datasets, after cross-validation of the four different numbers of hidden nodes, one network structure was chosen for each of the three learning methods:

1) Std_Backpropagation with three hidden nodes
2) BackpropChunk with 100 hidden nodes
3) SCG with 50 hidden nodes

Similar to the EqualFiltered datasets, plotting the positive and negative score distributions for the training set protein pairs after training the full networks (shown in Figure 3.10, below) indicated that each of the networks was training successfully.

#### 3.3.1.5 Blind Test of the EqualFam Dataset

To compare the performance of the three neural network predictors trained on the EqualFam dataset, each method was tested again against a blind test set with 1000 positive and 1000 negative pairs with SCOP family classifications not seen in any of the pairs in the training set. Figure 3.11 shows the full ROC curves and Table 3.2 provides the AUC values for each network.

Comparing the EqualFam blind test set performance to that of the EqualFiltered blind test set ROC plot (Figure 3.11) and AUC values (Table 3.2), there is no significant difference between each method's prediction capability for each of the method. This lack of considerable variation suggests that the superfamily composition of the training

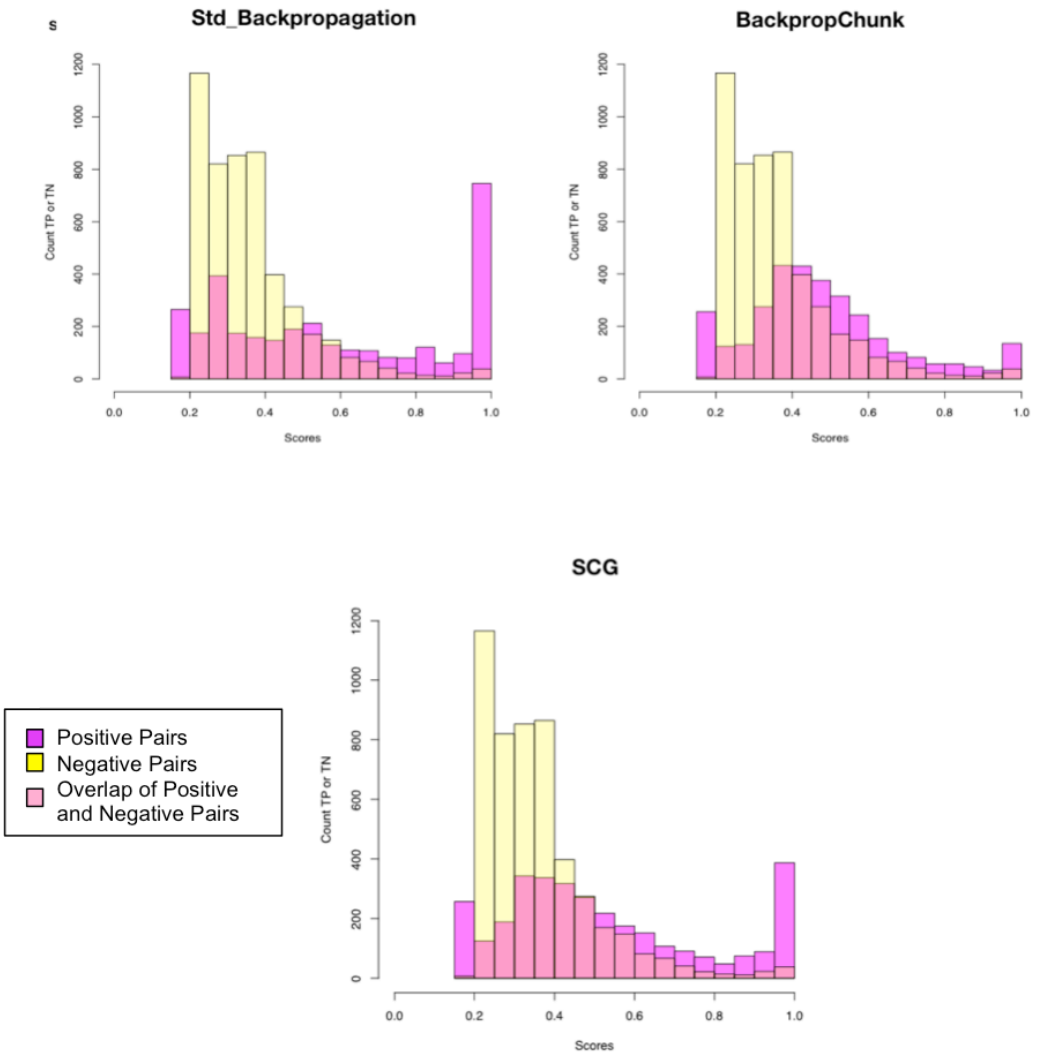


**Figure 3.10: Distribution of the scores assigned for the training set during training the Std_Backpropagation, BackpropChunk and SCG network combinations.** The three histograms of the distributions of scores assigned to the positive (pink) and negative (yellow) protein pairs are compared for scores assigned to the training examples during training of the full Std_Backpropagation, BackpropChunk and SCG with the EqualFam dataset. The distributions of scores for positive and negative pairs show no significant difference for each learning method (KS-test: p-value = $2.2e^{-16}$ for all methods) suggesting that the neural networks have successfully learned to discriminate between the examples.

set has had little to no bearing on the ability of the neural network to learn and characterise new data.

| Method | Area Under the Curve (AUC) EqualFiltered | Area Under the Curve (AUC) EqualFam | Statistical Difference between AUC (p-value) |
|---|---|---|---|
| SCG | 0.775 | 0.776 | 0.9452 |
| Std_Backpropagation | 0.775 | 0.813 | 0.0364 |
| BackpropChunk | 0.748 | 0.744 | 0.7538 |

**Table 3.2: Areas under the curve (AUC) for predictions in the EqualFam blind test set.** AUC values and p-values (Delong's test for two ROC curves) were calculated by the pROC package in R (Robin *et al.*, 2011) for the ROC curves for the predictions in the blind test set with the neural networks trained on the SCG, BackpropChunk and Std_Backpropagation learning methods.

As a result, the EqualFiltered dataset will remain the training dataset for the raw scores method for the SNNS PIPs predictor. While the networks trained with the SCG and Std_Backpropagation learning methods perform comparably in the blind tests (with the Std_Backpropagation method only slightly better), the SCG method performed more consistently throughout each stage of training and testing analysis. During initial cross-validation to select the optimal number of hidden nodes for each of the learning method networks, the SSE curve for the Std_Backpropagation network was not smooth throughout, particularly when compared to the smoothness of the SCG curve. As a result, the SCG learning method with a network structure of six input, 12 hidden and one output node was selected as the prediction method for the full protein pair set.

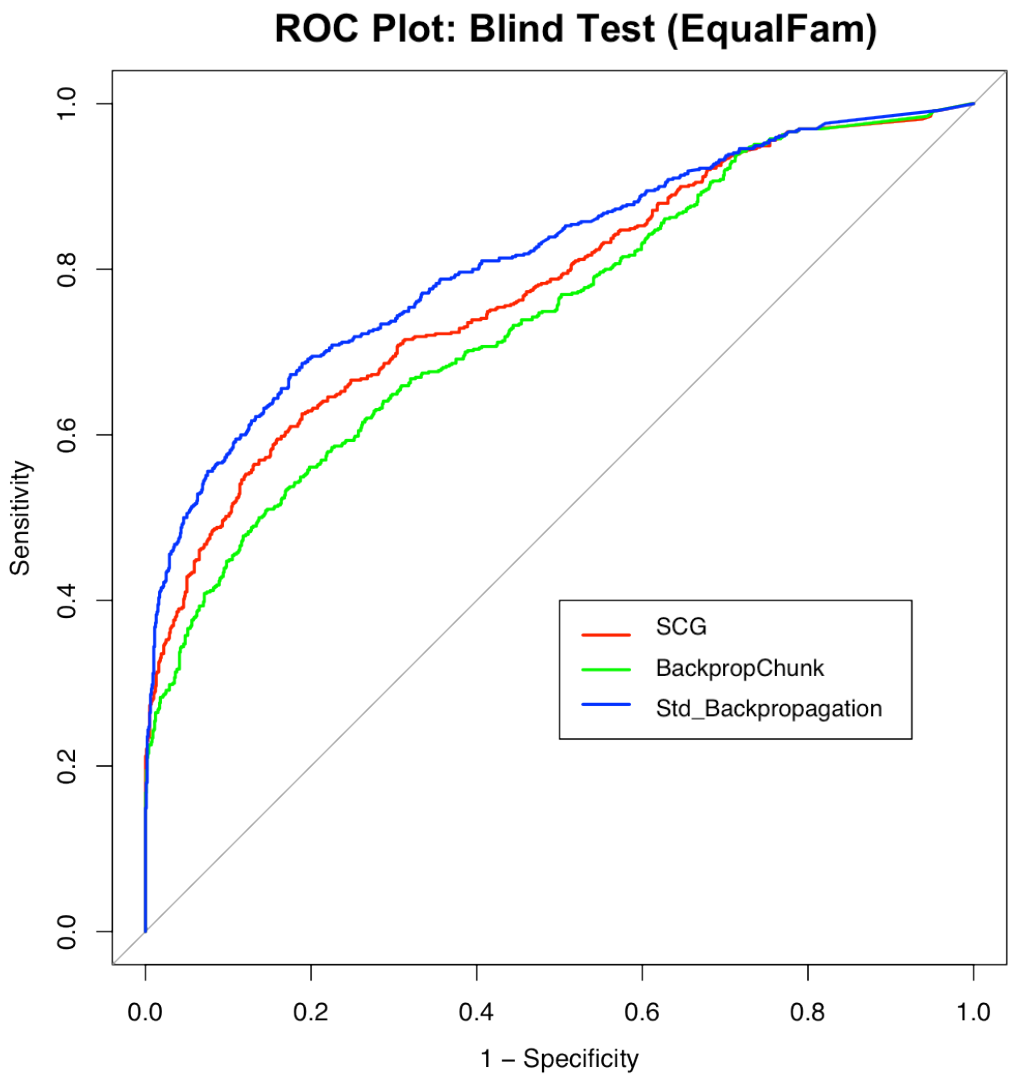


**Figure 3.11: ROC plot for predictions in the EqualFam blind test.** ROC curves of the prediction results for the blind test set by the networks trained on the EqualFam dataset with the SCG (red), BackpropChunk (green) and Std_Backpropagation (blue) learning methods are compared above. Curves were drawn by the pROC package in R (Robin *et al.*, 2011).

### 3.3.2 The Likelihood Ratios Method

An additional method of presenting data to the neural network, where each input value represented the likelihood ratio calculated by Bayesian PIPs for each of the six modules,

was also attempted. With no observed, negative effect of the superfamily composition of the training and testing datasets, the EqualFiltered dataset was kept as the training and testing set. As in the raw scores method, the optimal network structure for each of the three learning methods was first considered and resulted in the following combinations for training the full networks:

1) Std_Backpropagation with three hidden nodes
2) BackpropChunk with 100 hidden nodes
3) SCG with three hidden nodes.

The histogram distributions of the scores assigned to the training set examples during training, shown in Figure 3.12, showed that while each of the methods assigned positive pairs higher scores, there were also a large number of positives assigned low scores, particularly when compared to the raw scores method.

In order to assess if the likelihood ratio method was capable of predicting more accurately than the raw scores method, output scores were obtained for the 1000 positive and 1000 negative pairs in the EqualFiltered blind test set. Figure 3.13 compares the ROC100 curves for the likelihood ratios method with the curves for the raw scores method when trained and tested on the same dataset.

As the plot below indicates, the raw scores (red, green and blue) and likelihood ratios methods (orange, yellow and cyan) showed no significant difference in prediction accuracy on the EqualFiltered blind test for the SCG method (Z = -0.4443, p-value =

0.6568 from DeLong's test for two correlated ROC curves, as calculated by the pROC package in R (Robin *et al.*, 2011)).

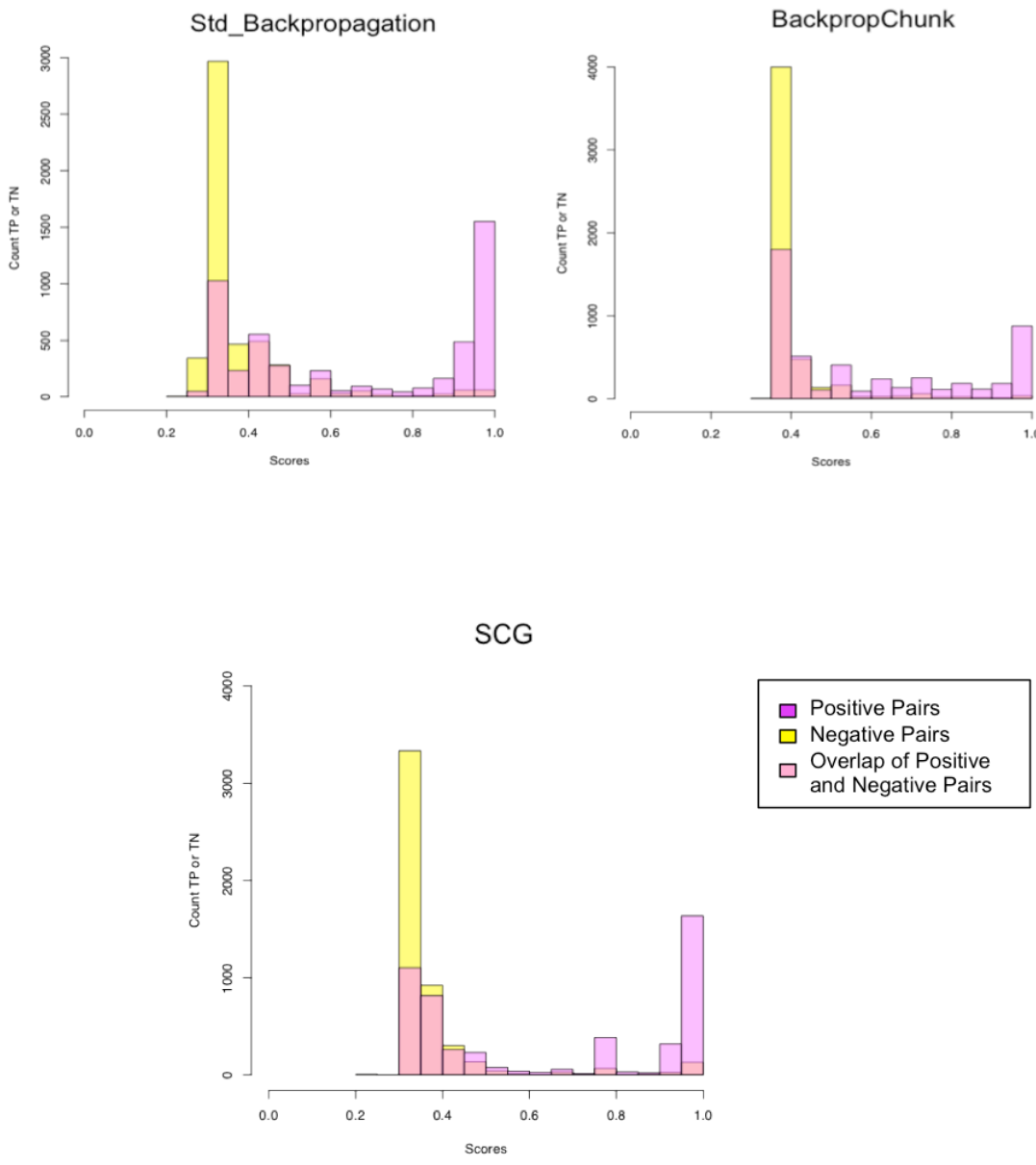


**Figure 3.12: Histograms of distributions of scores assigned to the EqualFiltered Dataset training set during training with the LR scores method.** The three histogram distributions of final output scores assigned to the positive (pink) and negative (yellow) training set examples during training of the Std_Backpropagation, BackpropChunk and SCG learning methods on the EqualFiltered dataset with the likelihood ratios method are compared above.

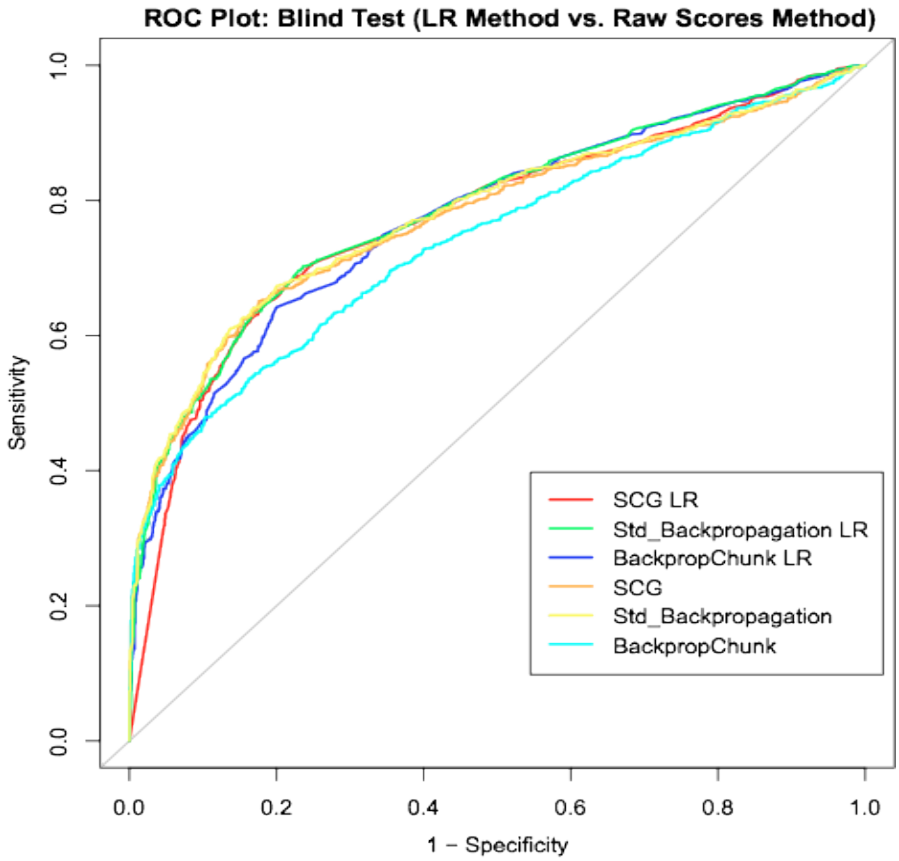


**Figure 3.13: ROC plot comparing predictions in the EqualFiltered blind test set for the raw scores and likelihood ratios methods.** The ROC curves calculated for predictions in the EqualFiltered blind test set for the SCG (red, AUC = 0.775), BackpropChunk (green, AUC = 0.739) and Std_Backpropagation (blue, AUC = 0.780) neural networks from the raw scores method are compared with the ROC curves for predictions in the EqualFiltered blind test set from the SCG (orange, AUC = 0.771), BackpropChunk (cyan, AUC = 0.772) and Std_Backpropagation (yellow, AUC = 0.784) networks from the likelihood ratios method. Plots were constructed with the pROC package in R (Robin *et al.*, 2011).

In order to examine if this lack of difference was due to the small size of the test sample set, networks trained with the SCG learning method on the raw scores and likelihood ratios input data were compared through a larger blind test set with 6523 positive and 6523 negative pairs.

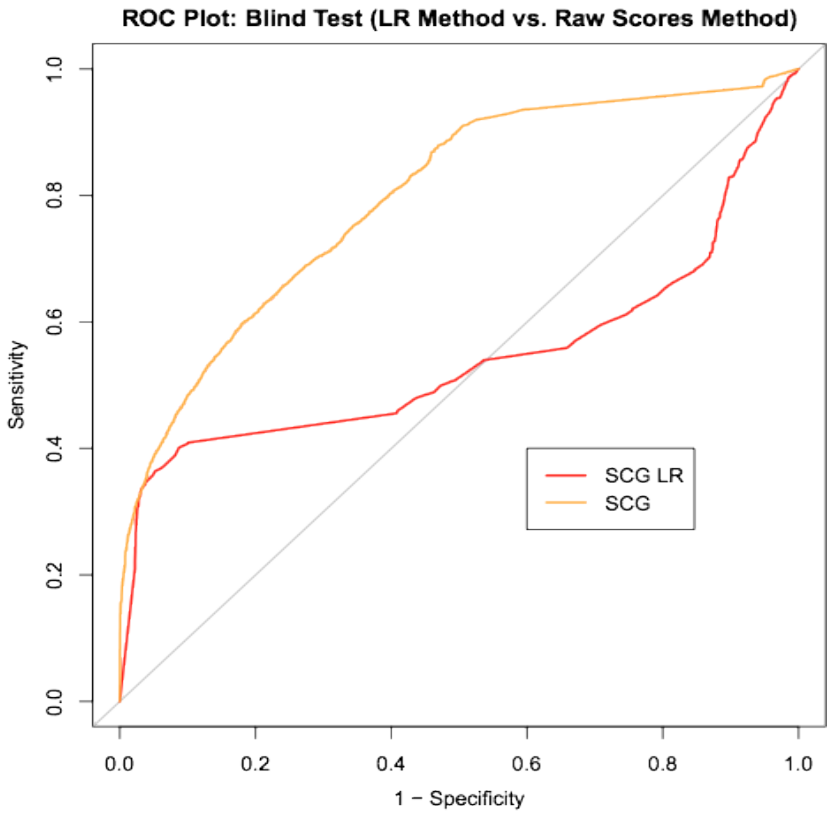


**Figure 3.14: Full ROC curve comparing the performance of the raw scores and likelihood ratios methods on a larger blind test set.** The performance of the neural network predictors trained with the SCG learning method on the raw scores and likelihood ratios input data are plotted as full ROC curves for a larger blind test set with 6523 positives and 6523 negatives. Of the two methods, the raw scores predictor (orange, AUC = 0.795) predicts with a consistency similar to its performance on the smaller test set (AUC = 0.771, Figure 3.12, above), while the likelihood ratios predictor (orange, AUC = 0.5424) shows a considerable decrease in accuracy (p-value = 2.2e-16, D=19.615, def=3133.9). Plots and values were calculated by the R package pROC (Robin *et al.*, 2011).

As the full ROC curve in Figure 3.14 indicates, increasing the size of the blind test set showed a substantial decrease in prediction accuracy for the likelihood ratios method (red) versus the raw scores method (orange). As a result, the likelihood ratios method was not pursued further.

### 3.3.2 Prediction of the Entire Set of Protein-Protein Interactions

A total of 702,589,781 protein pairs were included in the prediction set for the final neural network predictor. Table 3.3, below, provides the number of predictions above six potential thresholds: 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9.

| Threshold | Number of Predictions |
|---|---|
| 0.4 | 6,828,595 |
| 0.5 | 1,999,770 |
| 0.6 | 806,804 |
| 0.7 | 517,490 |
| 0.8 | 393,053 |
| 0.9 | 222,593 |

**Table 3.3: Number of pairs with predictions scores above cut-off thresholds.**

#### 3.3.2.1 Selection of a Cut-Off Threshold for Prediction

The final output scores from the neural network predictor range from 0.0 to 1.0, where 0.0 indicates a strong negative and 1.0 a strong positive result. Within the linear range, the value of the output prediction score corresponds to the strength of the validity of the prediction. In order to select a cut-off for what should be considered an interaction and not an interaction, the histogram of score distributions from training the network were considered again (see Figure 3.15, below) and the breakdown of true and false positives assessed at different potential cut-offs.

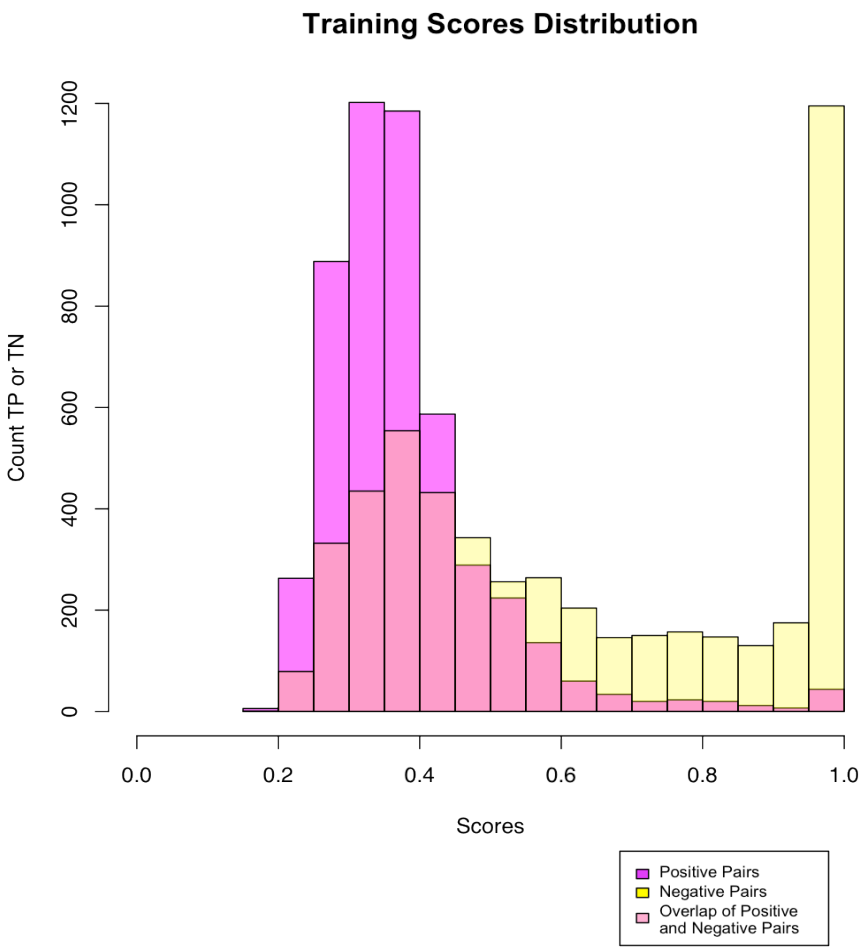


**Figure 3.15: Distribution of positive and negative scores assigned during training of with the SCG learning method on the EqualFiltered dataset.** The distribution of scores for the positive (pink) and negative (yellow) examples assigned during training of the neural network with the SCG learning method with 50 hidden nodes is significantly different (KS-test: D = 0.417, p-value = 2.2 e-16).

Logically, the most intuitive cut-off would be halfway between the minimum and maximum scores (i.e. 0.5). Considering the distribution of the score assignments to the positive and negative examples during training, 0.5 is a reasonable cut-off for minimising the number of false positive results returned without compromising the number of true positive predictions. The MCC, TPR and FPR for the training set at thresholds between 0.4 and 0.8 were also examined and are provided in Table 3.4, below.

| Threshold | MCC | TPR | FPR |
|---|---|---|---|
| 0.4 | 0.464 | 70.3% | 24.1% |
| 0.5 | 0.503 | 57.0% | 9.6% |
| 0.6 | 0.505 | 48.5% | 4.1% |
| 0.7 | 0.479 | 42.7% | 2.6% |
| 0.8 | 0.453 | 38.1% | 1.9% |
| 0.9 | 0.414 | 31.7% | 1.1% |

**Table 3.4: Matthew's Correlation Coefficient, TPR and FPR for results from training the SCG neural network.** Matthew's Correlation Coefficient (MCC), the True Positive Rate (TPR) and False Positive Rate (FPR) was calculated at thresholds of 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9 for the SCG neural network according to the predicted results on the training set after training the full predictor.

The difference between the values above at the 0.4, 0.5 and 0.6 thresholds was slight, with MCC values of 0.464, 0.503 and 0.505, respectively. TPR values (70.3%, 57.0% and 48.5%, respectively) and FPR values (24.1%, 9.6% and 4.1%, respectively) varied slightly more.

As an additional consideration, Figure 3.16 plots the distribution of the scores assigned to a larger blind test with 6523 positive and 6523 negative blind test set examples. While this histogram alone does not suggest a clear cut-off for positive and negative predictions, it does indicate that the predictor assigned the majority of the negative pairs very low scores (< 0.2). While there are also positive pairs assigned lower scores, the low number of negative pairs above 0.4 suggests that picking any cut-off in that mid-range will be discriminatory enough to minimise the number of false positive predictions.

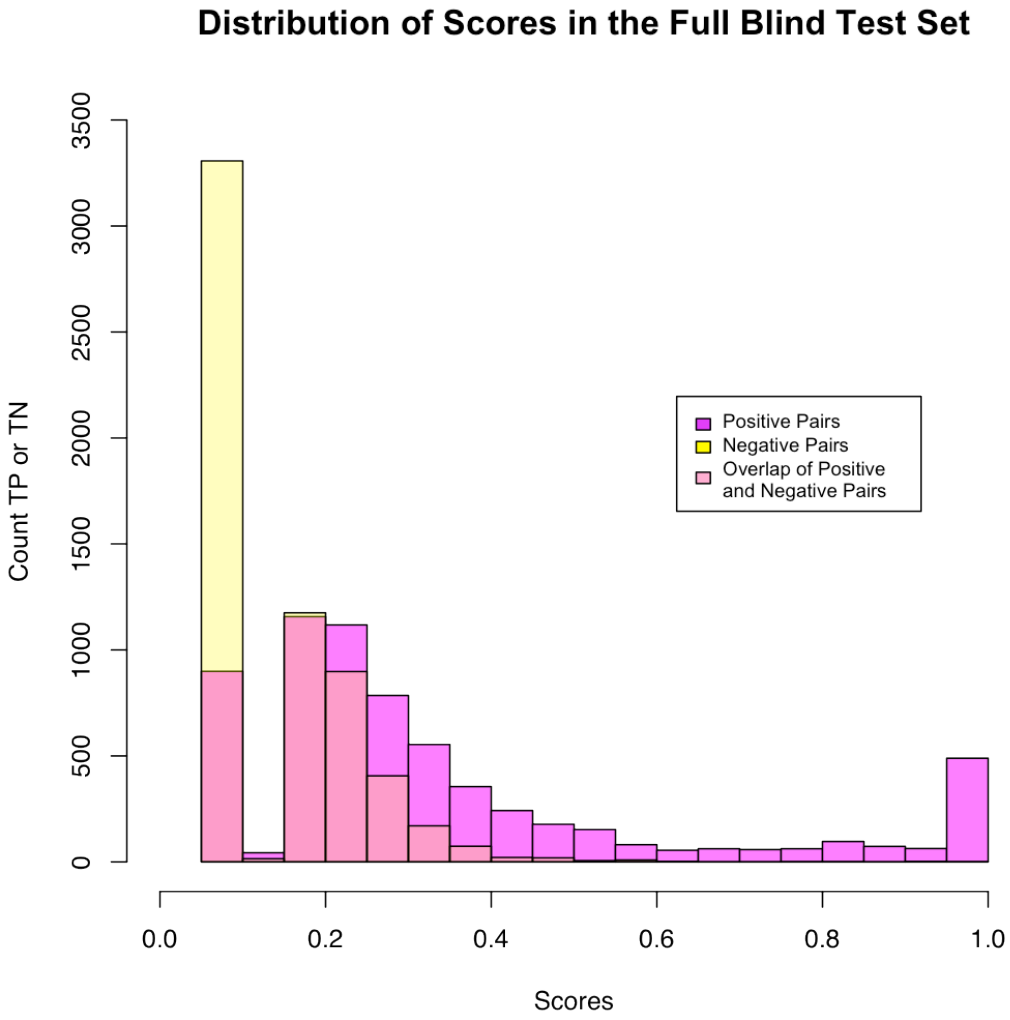


**Figure 3.16: Histogram of distribution of scores for predictions in the full blind test set.** The distribution of positive (pink) and negative (yellow) scores as calculated by the network trained on the EqualFiltered dataset with the SCG learning method are shown.

As an additional measure, Figure 3.17 plots the accuracy (the proportion of correct predictions, red) and precision (the proportion of true positive predictions in the set of all positive predictions, blue) at cut-off thresholds for output scores between 0.0 to 1.0. As expected, as the cut-off threshold was increased, the accuracy of prediction increased (i.e. low-scoring negatives were predicted as negative and high-scoring

positives were predicted as positive). Conversely, as the cut-off threshold was increased, the precision decreased (i.e. the highest cut-offs, only the highest-scoring positives were predicted as positive, at the expense of excluding the low- or mid-scoring positives).

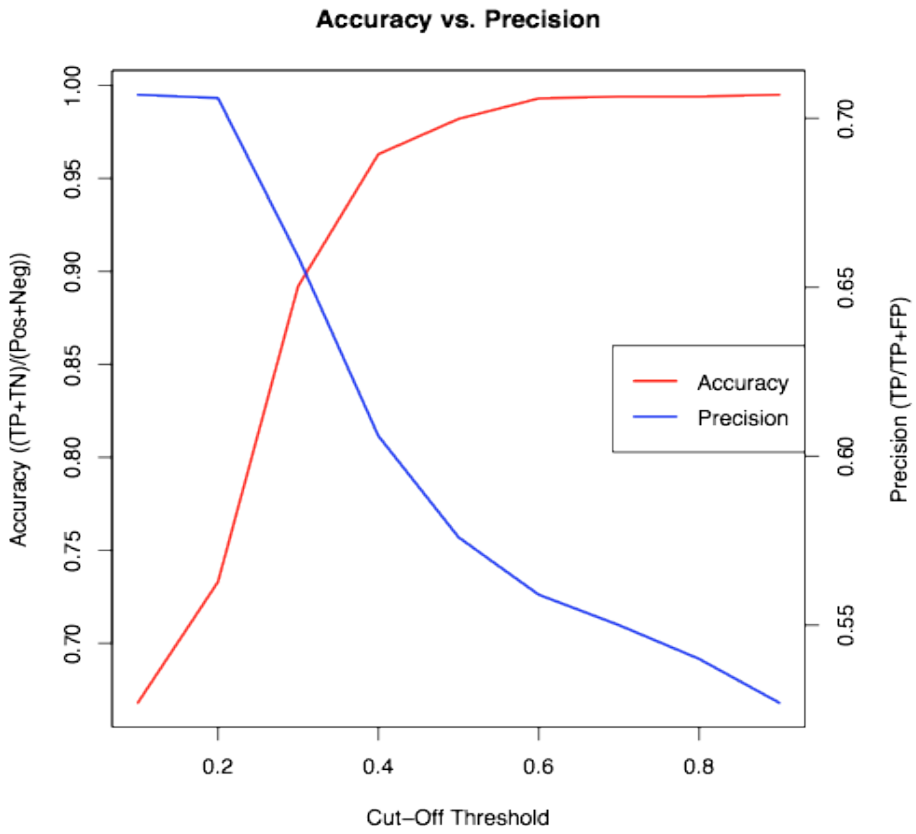


**Figure 3.17: Accuracy vs. Precision plot for SCG output scores in the full blind test set.** The accuracy (left-hand y-axis and red line) of predictions made on the full blind test set (6523 positives and 6523 negatives) was calculated by dividing the sum of true positives and true negatives by the total number of positives and negatives for output scores between 0.0 and 1.0 (x-axis). The precision (right-hand y-axis and blue line) of predictions was calculated by dividing the number of true positives by the total number of true positives and false positives at output scores between 0.0 and 1.0. As the cut-off threshold increases, the accuracy increases such that positive and negative pairs are predicted correctly. Conversely, the precision decreases such that at higher cut-off thresholds, fewer positives are predicted overall.

The point where the two plots cross (0.3, 0.9) represents the threshold at which the predictor was able to identify the highest number of interactions correctly without excluding positive predictions. When considered with the score distribution above

(Figure 3.14), a cut-off of 0.3 appears a reasonable prediction threshold for the selected blind test set. However, this blind test set only represents a small sampling of the set of possible protein pairings. With over six million interactions predicted across this set at a 0.4 cut-off, taking 0.3 as a threshold is likely far too non-specific for interactome-wide prediction.

Therefore, a cut-off threshold of 0.5 was selected for further stages of analysis. While this cut-off may exclude a number of positive interactions, the increased accuracy at thresholds above 0.5 suggests that the interactions predicted are more likely to be true positives. However, this cut-off should not be absolute, and pairs with prediction scores in the 0.4 to 0.6 range should not be immediately discounted in practical application of the predictor.

### 3.3.3 Incorporation of the Network Analysis

A strong aspect of the original Bayesian PIPs predictor is its incorporation of network analysis into its predictive framework. While each of the Expression, Orthology and Combined modules offer contributory evidence toward prediction, it is the Transitive, Cluster and TransMCL modules that provide an extra level of predictive power by assessing the context of each pair within the predicted network, rather than as an independent entity. As this networking principle is a unique aspect of the PIPs framework, an attempt was made to include the same method of analysis as in the Transitive module as a second stage to the neural network predictor.

Two different initial predicted interaction networks (the '0.5 Network', with pairs with initial output scores above 0.5, and '0.7 Network', with pairs with initial output scores above 0.7) were considered. After constructing these networks, a transitive score was calculated for each pair of proteins, and the results were processed in two ways. First, transitive scores for each pair were considered 'as is', i.e. a number between 0.0 and 184.0 for the 0.5 Network and a number between 0.0 and 543.8 for the 0.7 Network. In the second method, the output score from the first neural network and the normalised transitive score were submitted as input to a second neural network that was trained on the same dataset as the original network to give a new output score between 0.0 and 1.0. The results for both methods for predictions on the EqualFiltered blind test set (i.e. with 1000 positives and 1000 negatives, as used in Section 3.3.1.4 and Figure 3.8, above) were plotted as ROC curves in Figure 3.18 along with the SCG (red), BackpropChunk (blue) and Std_Backpropagtion (green) methods trained on the EqualFiltered dataset.

The first scoring method of taking the transitive scores as they were was unsuccessful at discriminating between positive and negative examples, with both the 0.5 Network (pink) and 0.7 Network (orange) ROC curves for the blind test set barely above random (grey line bisecting the plots). However, both the 0.5 (purple) and the 0.7 Networks (cyan) in two-stage neural network method performed comparably to the one-stage neural networks without the transitive analysis component. Additionally, there was a significant difference (p-value = 4.87E-06, Delong's test for two ROC curves, as calculated by pROC (Robin *et al.*, 2011)) between the two cut-off thresholds chosen to construct the initial predicted network.

Interestingly, while it was expected that the more stringent 0.7 cut-off threshold (cyan) would be the better of the two, the 0.5 Network (purple) predicted more accurately on the blind test. It is, therefore, likely that although increasing the threshold has assembled a smaller, strong network of potential interactors, the size and sparsity of the network has limited the amount of new information that the transitive analysis component can add. Additionally, at 1,999,770 pairs, the size of the 0.5 Network more closely mirrors the size of the initial predicted interaction network considered by PIPs (1,190,825), which was optimised when originally designed, than the 0.7 Network, which at 517,490 pairs is a quarter of the size. Conversely, decreasing the cut-off to 0.4 will most likely have the opposite effect of increasing the initial network to three-and-a-half times the size of the 0.5 Network (6,828,595 pairs). With a drastic increase in the false positive rate between the 0.4 (24.1%) and 0.5 (9.6%) cut-off thresholds, the resulting network is likely to include too many false positive predictions for the transitive analysis to be valuable.

Comparing the ROC curves from the 0.5 Network (the best of the two two-stage neural networks tried) with the SCG one-stage neural network showed no significant difference between the methods (p-value = 0.070). However, further examination of the scores assigned to the training set during training, plotted as a histogram in Figure 3.19, shows that there is a significant difference in the distribution of output scores for known positives and negatives between the one-stage neural network and the two-stage transitive neural network.

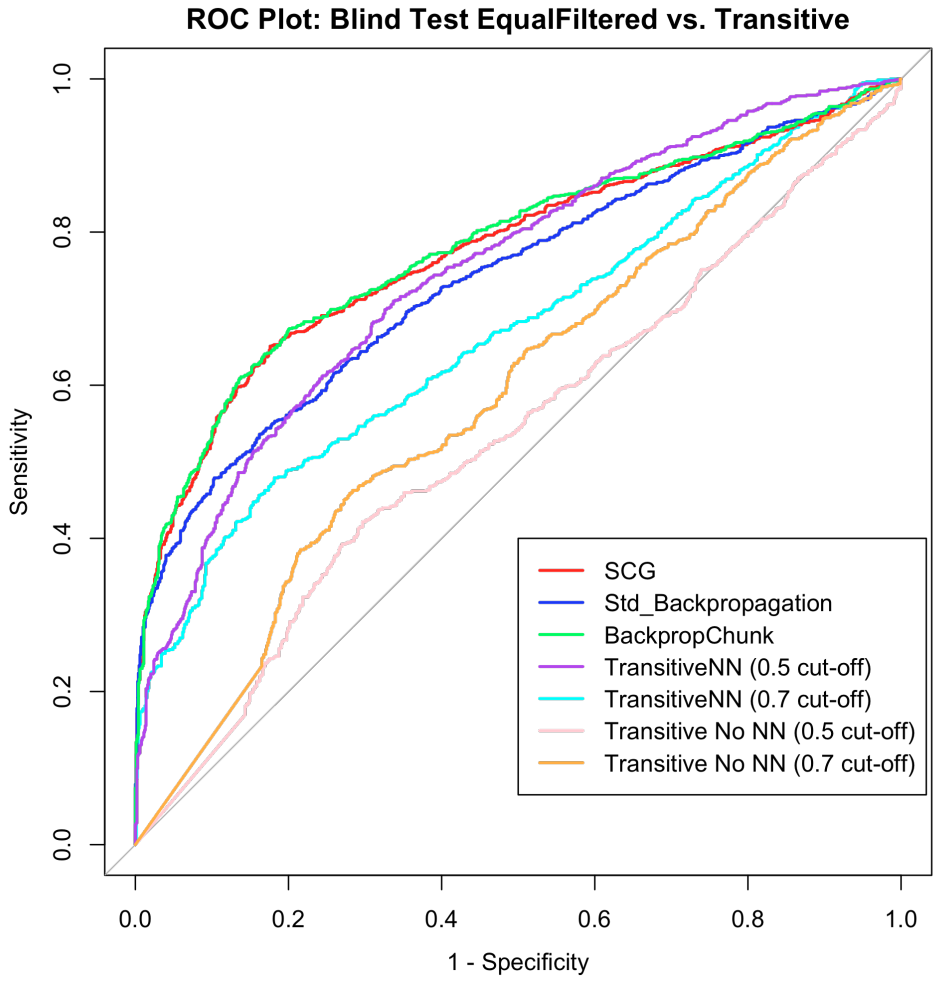


**Figure 3.18: ROC plot comparing predictions in the EqualFiltered blind test for one- versus two-stage predictors.** ROC curves for the two methods of incorporating network analysis into the neural network PIPs framework with the transitive scores alone predictions for the 0.5 Network (pink, AUC=0.535) and 0.7 Network (orange, AUC=0.589) and with the second neural network step for the 0.5 Network (purple, AUC=0.748) and 0.7 Network (cyan, AUC=0.668) assessing accuracy of predictions in the EqualFiltered blind test set. As a comparison, the ROC curves for the outcomes of the three previous predictors trained on the EqualFiltered dataset with the SCG (red, AUC = 0.775), BackpropChunk (green, AUC = 0.739) and Std_Backpropagation (blue, AUC = 0.780) learning methods on the same blind test set are also shown. While an unpaired T-test comparison of the SCG and Transitive method ROC curves indicates no significant difference (D = 1.814, dof = 3988.667, p-value = 0.070), the lower ROC profiles and AUC value for the Transitive method suggests it predicts less accurately than the one-stage neural network. Curves and calculations were computed with the R package pROC (Robin *et al.*, 2011).

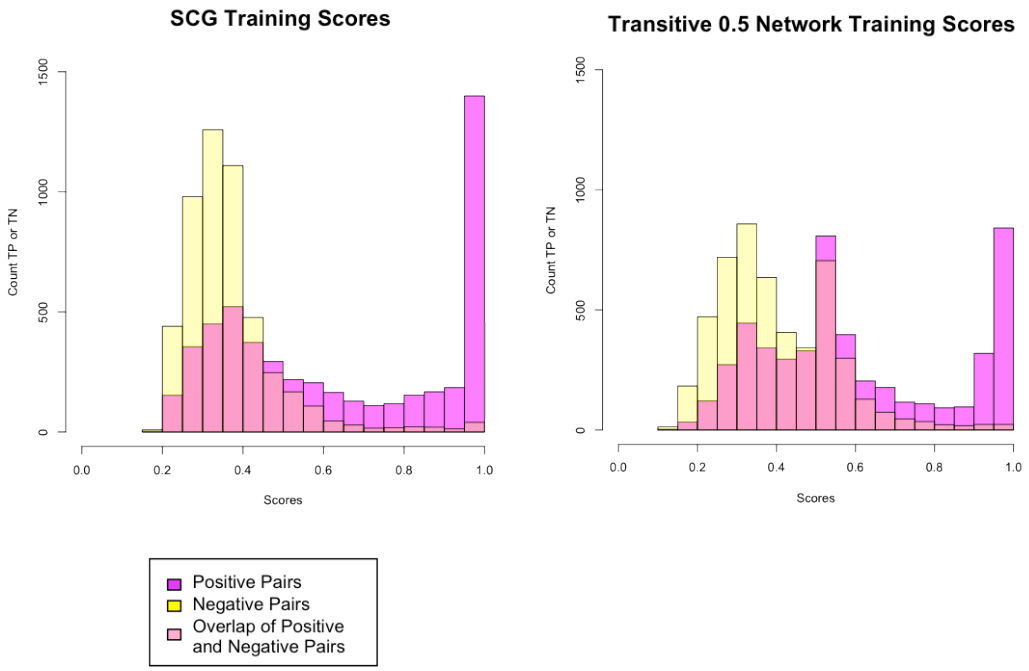


**Figure 3.19: Comparison of distributions scores assigned to positive and negative training pairs during training the EqualFiltered SCG method with and without the Transitive NN.** The distribution of scores assigned to the positive (pink) and negative (yellow) pairs in the EqualFiltered training dataset during training without the transitive analysis component (left), with the 0.5 Network and transitive analysis component (right) are plotted above. While both the SCG learning method without the transitive analysis and the 0.5 Network with the transitive analysis appear to have assigned the majority of positives with high scores and negatives with low scores, there is a significant difference between both the positive (Wilcoxon T-Test p-value = $1.38e^{-14}$, W=13611098) and negative (p-value = $2.2e^{-16}$, W=10809970) score distributions between the methods.

Additionally, a large number of positives have been assigned low scores, particularly when compared to network trained on the same dataset without the transitive neural network.

Overall, incorporation of the transitive network analysis into the neural network predictor has failed to significantly increase predictive capability. Therefore, at this the time, the Transitive module of Bayesian PIPs will not be considered for further analysis.

### 3.3.4 Final SNNS PIPs Predictor

The final SNNS PIPs predictor, now called PIP'NN, consists of a neural network trained on a dataset of 5000 positive and 5000 negative pairs with the SCG learning method on a network structure of six input, 12 hidden and one output nodes. The raw scores supplied to the original Bayesian PIPs predictor for the Expression and Combined modules are provided as five input values with the likelihood ratio calculated for the Orthology module and are normalised to values between 0.0 and 1.0.

## 3.4 Discussion

Developing the final PIP'NN predictor proved to be a multi-step process requiring the testing of numerous combinations of parameters, precise selection of the training set and establishment of the unbiased blind testing set. While the first dataset selected, the EqualLarge dataset, failed to provide each of the neural networks with enough input evidence to train successfully, filtering the dataset to include only protein pairs with data available from multiple sources of evidence (the EqualFiltered dataset) did enable the network to learn. While this filtering appeared crucial to getting the networks to train, it could also be detrimental to how the networks will perform when presented with input patterns from protein pairs that lack known data for the majority of the six sources of evidence provided to the predictor. However, testing the prediction capabilities of the three networks with the blind set for the EqualFiltered datasets, which was not filtered, indicated that each of the three neural networks were still able to classify both the known positive and known negative pairs accurately.

The initial results for the three neural networks trained on the EqualFiltered datasets and the raw scores method of data presentation showed a promising prediction accuracy, and it was necessary to scrutinise how the neural network was being trained and tested to ensure that it was not learning and being assessed on too similar datasets. However, designing the new EqualFam dataset and blind test set, where protein pairs were separated into training and testing groups based on their superfamily classifications, showed only a negligible difference between the performance of the three neural network predictors. Together, the solid performance of all of the neural network predictors trained on both the EqualFiltered and EqualFam datasets suggests that overall, the neural network method is able to learn to correctly predict interactions between proteins pairs similar and different to those it has and has not seen during training.

Unfortunately, the second method of data presentation to the network, the likelihood ratios method, was less successful than the raw scores method at distinguishing positive and negative interactions. One possible explanation for this decrease could be attributed to the data normalisation aspect; while standardising the likelihood ratios to between 0.0 and 1.0 does linearise them onto a smaller scale, it might cause the lower likelihood ratios to become too low to be considered properly by the network. In the future, one option could be to test the likelihood ratios method with $\log_{10}$ normalisation of the ratios instead to see if the different scaling has any effect.

Additionally, while it was hoped that incorporating the principles of network analysis in use with the Transitive module in Bayesian PIPs would boost the performance of the predictor, both handling the transitive scores as prediction scores on their own and

including them via a second neural network did not improve prediction accuracy. Interestingly, between the two different sized initial predicted networks, the lower 0.5 cut-off and larger initial network performed better of the two. However, after closer examination of the network construction, it is likely that the smaller size of the 0.7 Network decreased the information that could be gained from analysis, while taking a lower cut-off threshold of 0.4 would create too large of a network to be valuable. Regardless, the two-stage neural network predictor with the 0.5 Network failed to significantly increase prediction accuracy, and the Transitive network analysis was not incorporated into the final PIP'NN predictor.

One downside to the neural network method is the loss of the ability to easily view how each individual component (i.e. module in PIPs) has contributed to the final prediction score outcome. While the individual raw scores input into the neural network can still be viewed, how this information is processed and weighted by the network remains lost in the metaphorical 'black box' of the prediction method. As such, presentation of prediction results could include a breakdown of the supplied evidence to further support or discredit predictions.

Additionally, while there appears to be a clear distinction between the scores assigned to the known positive and negative pairs in the training and testing datasets, there is not a clear cut-off for what definitively denotes an 'Interaction' or 'No Interaction'. As a result, final results for the predictor should not be limited to a hard-fast cut-off of 0.5 but should also consider those pairs with scores in the 0.4 to 0.6 range, with the understanding that the predictions of particular interest are those with scores in the highest score range.

Finally, a neural network structure with two output nodes instead of one was not yet attempted. While one output node allows a straightforward assignment of one output score corresponding to the final prediction, two output nodes, where one value represented 'Predicted Interaction' and one value represented 'No Predicted Interaction', would provide a measure of confidence. For example, a large difference between the output scores for the 'interaction' and 'no interaction' scores (i.e. if 'Predicted Interaction' were 0.9 and 'No Predicted Interaction' were 0.1) would indicate that the pair was predicted strongly as interacting and weakly as not interacting. Likewise, a negligible difference (i.e. if 'Predicted Interaction' were 0.61 and 'No Predicted Interaction' were 0.49) would suggest that the neural network was not able to classify the pair definitively one way or the other. Therefore, future development of PIP'NN should investigate this option to add a second level of discrimination and make predictions stronger.

## 3.5 Conclusions

1) The network structure and number of hidden nodes required for the three learning methods (Std_Backpropagation, BackpropChunk and SCG) varies between the method and the training dataset.

2) To successfully train each of the learning method-hidden node network combinations, it was was necessary to filter the training dataset to include protein pairs with available data for at least half of the sources of input evidence.

3) The networks trained on the EqualFiltered and EqualFam datasets were able to correctly classify the known positive and negative interactions in blind test sets

composed of protein pairs similar, but not identical, to those included in the training dataset and in sets composed of protein pairs structurally and physiologically different from those in the training dataset.

4) A second method of data presentation to the neural network, as normalised likelihood ratios as calculated by the Bayesian PIPs predictor, failed to improve performance.

5) Incorporation of the network analysis principle in the Transitive module in Bayesian PIPs in two different methods did not improve prediction accuracy from the one-stage neural network predictor with only data from the Expression, Orthology and Combined modules. Of the two different sized initial predicted networks, the two predictors based on the less stringent and larger network assembled with pairs scoring above 0.5 performed better than the two based on the smaller, more stringent network with pairs scoring above 0.7.

6) The final SNNS predictor has been trained on the EqualFiltered dataset with the Scaled Conjugate Gradient learning method with one input node, 12 hidden nodes and one output node.

7) An output score of 0.5 was selected as the final cut-off, with outcomes for pairs with scores above 0.5 considered 'interaction' and below 0.5 considered 'no interaction'. However, the large number of interactions predicted at this cut-off (approximately two million), suggests that a large number of pairs with mid-range scores are false positives. Therefore, predictions should be considered logically, with pairs with 0.4 to 0.6 output scores scrutinised manually for likeliness of interaction and those with the highest scores (0.7-1.0) considered the strongest.

8) Future development of PIP'NN should investigate neural networks with two instead of one output node.

# Chapter 4

## PIPs vs. PIP'NN: A Comparison of Predictive Capability

### Preface

---

This chapter compares the predictive capability of the Bayesian theory-based PIPs predictor with the new, neural network-based PIP'NN predictor. The predictors are compared against each other with several blind test sets of different sizes and compositions. Additionally, the top predictions from the PIPs predictor are compared with their outcomes from the PIP'NN predictor. Finally, both predictors are compared against other current human protein-protein interaction methods.

# 4.1 Introduction

Ultimately, the goal of re-engineering the PIPs framework with a neural network is to increase the number of true positive predictions with minimal increase in false positive results. While a fully accurate and consistent predictor will not exist until everything is known about every single cellular interaction, the evidence incorporated into machine learning techniques and how that information is handled can strongly influence the prediction quality. Despite the advantages for the naïve Bayesian theory as the underlying predictor framework, particularly that its output represents a quantifiable description of how likely an interaction is to occur rather than a discrete 'Interaction' or 'No Interaction' result, there is more merit in having a predictor with a consistently higher prediction accuracy. Additionally, with its enhanced learning capability, the neural network framework allows for much more flexibility in training the predictor by requiring smaller datasets and less time.

To assess the performance of the Bayesian PIPs predictor against the neural network PIP'NN predictor, outcomes for pairs in a blind test set composed of protein pairs not included in the training datasets of either predictor were first compared. Additionally, the highest scoring predictions in PIPs were compared with their prediction results from PIP'NN and the overlap between the sets of total predicted pairs for both predictors was determined. Finally, PIPs and PIP'NN have been compared and contrasted against other current predictors of human protein-protein interactions.

# 4.2 Methods

## 4.2.1 Blind Test Sets

First, the predictions were made by the PIPs and PIP'NN predictors for the 1000 positive and 1000 negative pairs in the EqualFiltered blind test set (see Chapter 3.2.3: Datasets). Additionally, predictions for pairs in a second blind test set with 5000 positives and 5000 negatives were also compared. Final likelihood ratios from PIPs and final output scores from PIP'NN were then plotted as ROC curves and ROC100 and areas under the curve (AUC) were calculated with the R package pROC (Robin *et al.*, 2011).

## 4.2.2 Comparison of Prediction Sets

To assess prediction concurrence between PIPs and PIP'NN, the top 50 predictions for each of the EOCT, EOCM and EOCZ methods in PIPs were obtained along with the PIP'NN results for each pair in the set. Additionally, the total set of interactions for both methods were also compared for overlapping and unique predictions. For PIPs, a cut-off of 1.0, after multiplying the final $LR_{EOCT}$, $LR_{EOCM}$ and $LR_{EOCZ}$ scores by the 1/1000 prior odds ratio, was selected, with pairs with scores above 1.0 labelled 'Interaction' and pairs with scores below labelled 'No Interaction'. For PIP'NN, a cut-off of 0.5 was chosen, with pairs above labelled 'Interaction' and pairs below labelled 'No Interaction'.

### 4.2.3 Comparison of PIPs and PIP'NN with Other Human Protein-Protein Interaction Prediction Methods

The main barrier to completing a objective comparison of prediction capabilities lies in the difference in positive and negative training datasets for each method. The current version of PIPs is trained and tested on a positive dataset consisting of 38,995 pairs of proteins included in the HPRD most recent release from April 2010, a subset of which is included in the PIP'NN positive dataset. Therefore, a starting dataset, downloaded from the PrePPI webserver (Zhang *et al.*, 2012, http://bhapp.c2b2.columbia.edu/PrePPI/downloads.html), that included 9811 interactions added to the HPRD between August 2010 and August 2011, was selected. This dataset was then filtered at four levels. First, all self-interactions and interactions already included in either the I2D, IntAct, BioGRID or DIP databases were removed, resulting in a total of 5178 interactions with 3759 unique proteins. Next, this smaller set was filtered such that no protein appeared in more than five interactions. The remaining 3270 pairs were filtered to 1659 interactions containing 3141 unique proteins. Finally, to generate final comparison set, the set of 1659 was filtered to 748 interactions that had evidence for at least one of the Expression, Orthology or Combined modules.

Five different methods were selected for comparison: STRING (Szklarczyk *et al.*, 2011), FunCoup (Alexeyenko *et al.*, 2012), IntNetDBv.1.0 (Xia, Dong & Han, 2006) and BIPS (Garcia-Garcia *et al.*, 2012), all of which consider gene context or orthology-based evidence, and PrePPI (Zhang *et al.*, 2012), a newly developed method that also considers structure. For each method, prediction data was downloaded, if possible, from the tool website or acquired via the group directly. Protein pairs in each dataset were then mapped to their UniprotAC accession identifiers via either a mapping file

provided by the service (i.e. STRING) or with the UniprotKB mapping tool. Data files were then parsed to extract protein pairs in the selected test set and their associated prediction scores or outcomes. Labels of 'Predicted Interaction' or 'No Predicted Interaction', provided in Table 4.1, were determined based on the specific prediction criteria of each tool as stated by the literature.

| Prediction Method | Prediction Criteria |
|---|---|
| PIPs | Final likelihood ratio > 1000.0 |
| PIP'NN | Final output score > 0.5 |
| STRING | Low confidence: score < 0.4; Mid confidence: score >= 0.4 and < 0.7; High confidence: score >= 0.7 |
| PrePPI | Final likelihood ratio > 600.0 |
| FunCoup | Final score > 4.0 |
| IntNetDB v. 1.0 | Final score > 6.0 |
| BIPS | Inclusion in dataset |

**Table 4.1: Scoring criteria for the interaction prediction methods compared.** Criteria for prediction classifications were collected from the most recent publication for each method.

As a second comparison, the number of interactions predicted by each of the methods above for 1211 pairs of proteins in the Negatome (Smialowski *et al.*, 2010), a manually curated database of interactions shown through structural studies not to interact, were obtained. Pairs were considered predicted as interacting according to the same thresholds as for the positive comparison (Table 4.1, above). In order to further analyse these results, full ROC and ROC50 curves were plotted for the three PIPs methods, PIP'NN and PrePPI taking the positive dataset as the 748 HPRD pairs described above

and the 1211 Negatome pairs.

# 4.3 Results

## 4.3.1 Blind Test Comparison of the PIPs and PIP'NN Predictors

As an initial blind test, prediction outcomes for the 1000 positive and 1000 negative pairs in the EqualFiltered blind test set were obtained for the EOCT (orange), EOCM (yellow) and EOCZ (cyan) methods of the PIPs predictors and the SCG method of the PIP'NN predictor. The full ROC curves for each method are plotted in Figure 4.1.

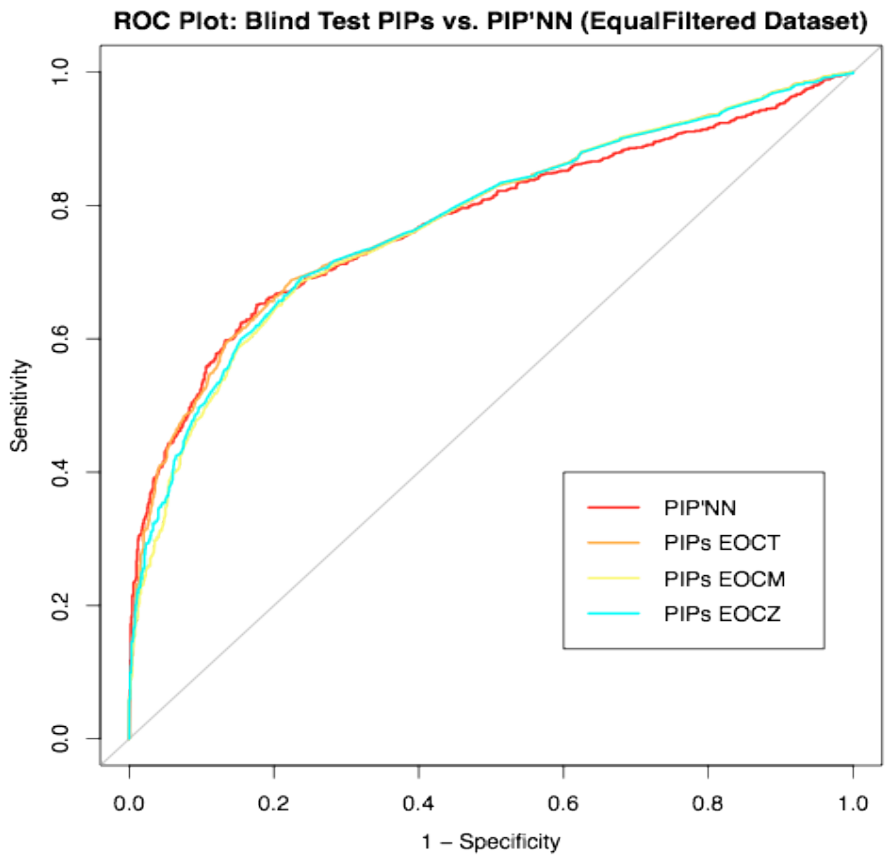


**Figure 4.1: ROC curves for predictions for pairs in the EqualFiltered blind test for the PIPs and PIP'NN predictors.** ROC curves for predictions from the PIP'NN (red) and PIPs EOCT (orange), EOCM (cyan) and EOCZ (yellow) methods on the 1000 positive and 1000 negative pairs in the EqualFiltered blind test set. Curves were constructed with the R package pROC (Robin *et al.*, 2011).

AUC values for the four curves are provided in Table 4.2, below. Interestingly, there is little difference between the PIP'NN predictor and the three PIPs predictors (with an unpaired T-test, D = 0.069, df = 3329.423, p-value = 0.945, as calculated by pROC in R (Robin *et al.*, 2011)). In order to investigate this similarity further, predictions for the PIPs EOCT, EOCM and EOCZ methods and the neural network predictor trained with the SCG learning method on the EqualFam dataset were computed.

| Method | EqualFiltered AUC | EqualFam AUC | Statistical Difference between AUC (p-value) |
|---|---|---|---|
| PIP'NN | 0.786 | 0.776 | 0.9452 (D = - 0.069, df = 3328.4) |
| PIPs: EOCT | 0.789 | 0.640 | 3.342E-13 (D = -7.315, df = 2776.7) |
| PIPs: EOCM | 0.774 | 0.640 | 1.496E-11 (D = -6.776, df = 2810.3) |
| PIPs: EOCZ | 0.775 | 0.640 | 2.827E-12 (D = -7.018, df = 2791.6) |

**Table 4.2: Comparison of areas under the curve (AUC) for the ROC curves for predictions in the EqualFiltered and EqualFam blind test sets.** Side-by-side comparison of the AUC values for the PIP'NN and PIPs ROC curves constructed for predictions for pairs in the EqualFiltered (column 2) and EqualFam (column 3) blind test sets and the p-value as calculated by Delong's test for two ROC curves. Values were calculated from the ROC plots in Figures 4.1 and 4.2, respectively, with the pROC package in R (Robin *et al.*, 2011).

As little difference was seen between the neural network predictors that were trained on the EqualFiltered and EqualFam datasets (i.e. not separated and separated by superfamily classifications, respectively, see Chapter 3.3.1.5: Blind Test of the EqualFam Dataset), testing the neural network predictor on this dataset should not affect its performance. Figure 4.2, below, shows the resulting ROC curves for predictions

from each of the four predictors. AUC values for both the EqualFiltered and EqualFam blind tests are given in Table 4.2.

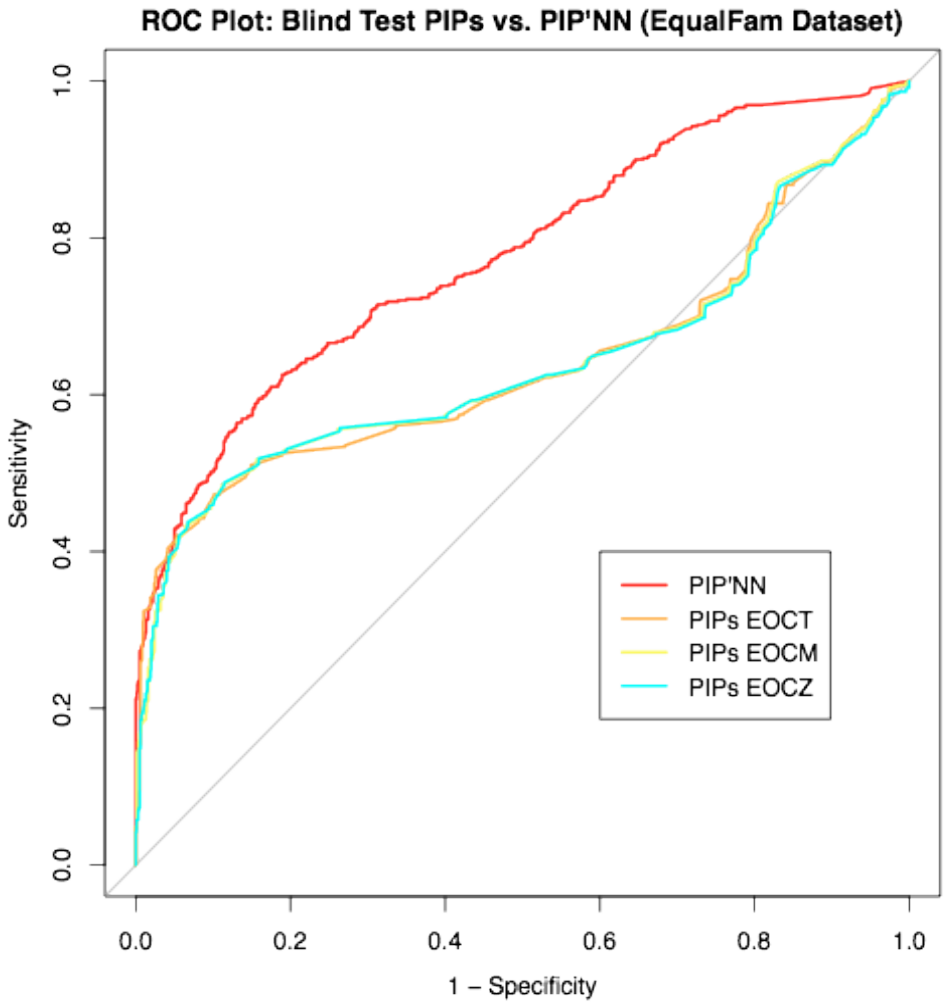


**Figure 4.2: ROC curves for predictions for pairs in the EqualFam blind test for the PIPs and PIP'NN predictors.** ROC curves for predictions from the PIP'NN (red) and PIPs EOCT (orange), EOCM (cyan) and EOCZ (yellow) methods on the 1000 positive and 1000 negative pairs in the EqualFam blind test set. Curves were constructed with the R package pROC (Robin *et al.*, 2011).

Surprisingly, while the accuracy of PIP'NN (red) was comparable with both blind test sets, the accuracy of the Bayesian PIPs predictors (orange, cyan and yellow), which should not have been affected by the set composition, decreased with the EqualFam blind test set.

One explanation for the varied performance of the Bayesian PIPs methods could be the size and composition of blind test set itself. With only 1000 positive and 1000 negative pairs, the set is only a small representation of the entire set of protein interaction predictions and the full set of data considered during training with the Bayesian method. With that much of a variation observed between selecting two different sets of 2000 pairs, it is likely that the accuracy of the Bayesian PIPs predictor is highly dependent upon the exact subset of proteins in the blind test set. Similarly, the PIP'NN might also have predicted well on the both blind test set without the performance being representative of its overall predictive capability.

As a result, a second, larger blind test set, with 5000 positive and 5000 negative examples, was constructed to provide a better assessment of each method's performance on a wider range of samples. Figure 4.3 plots the full ROC curves for PIP'NN and the three PIPs predictors for prediction on this larger test set. While the prediction accuracy of the PIP'NN predictor (red) remained consistent with the two smaller blind test sets, the accuracy of all three PIPs methods decreased with the larger set (orange, cyan and yellow).

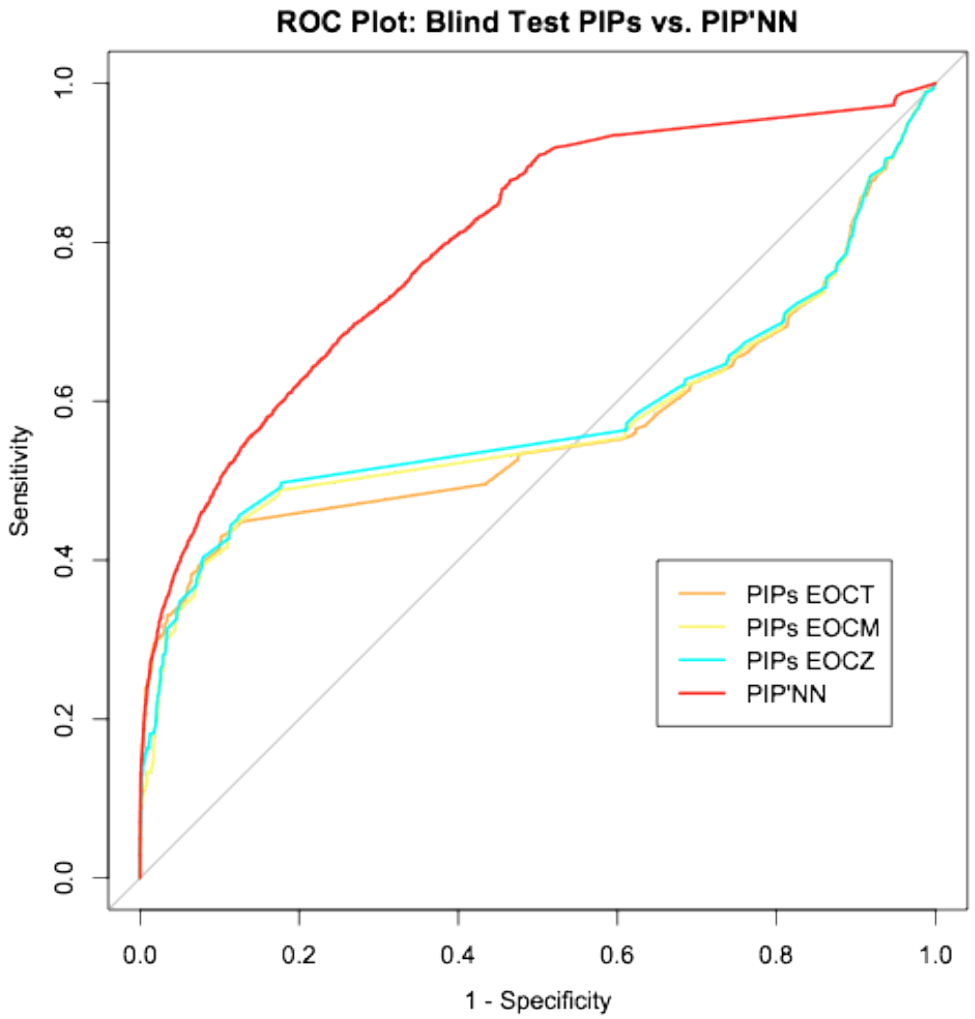


**Figure 4.3: ROC plot of the accuracy of predictions on the full blind test set for the PIP'NN and PIPs predictors.** Plot of the four curves corresponding to prediction accuracy of PIP'NN predictor (red) and PIPs EOCT (orange), EOCM (cyan) and EOCZ (yellow) predictors for 5000 positive and 5000 negative pairs in a blind test set. Plots were constructed with the R package pROC (Robin *et al.*, 2011).

As the sharp decrease in accuracy of the three PIPs predictors is surprising, the distribution of PIPs scores (shown as $\log_{10}$(final likelihood ratio) before adjustment for the 1/1000 prior odds ratio) within the blind test was plotted in Figure 4.4.

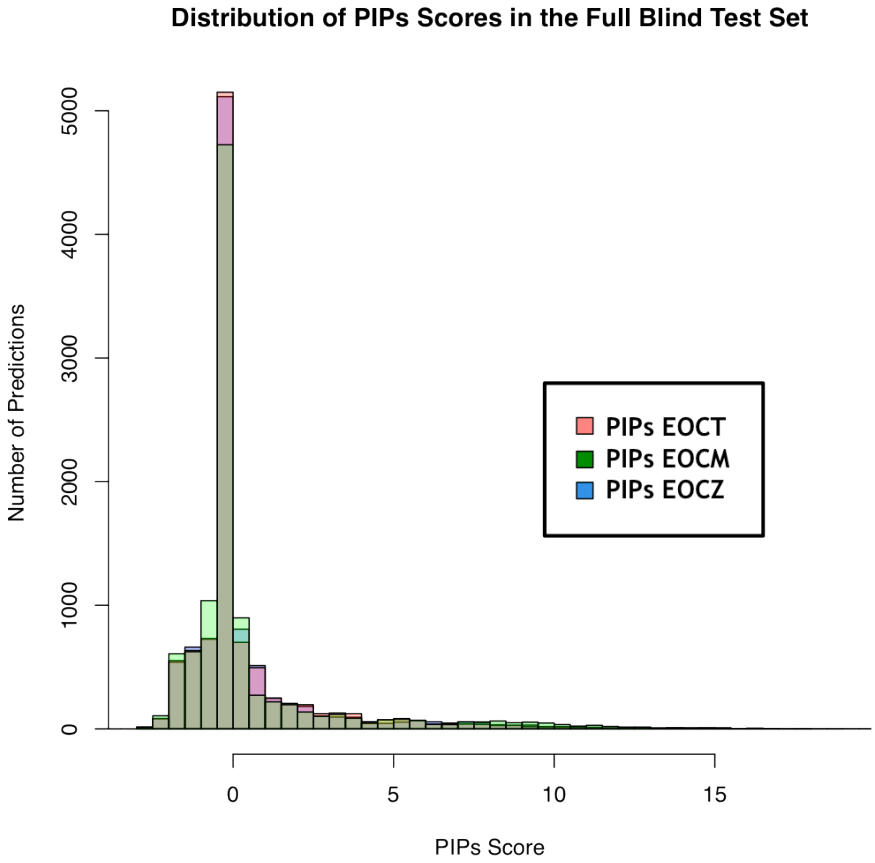


**Figure 4.4: Distribution of PIPs final scores in the full blind test set.** The distributions of the $\log_{10}$ of the final scores for the PIPs EOCT (red), EOCM (blue) and EOCZ (green) methods are plotted above. While there is a slight variation between the score distributions for each method, the majority of final scores in all fall around 0, or $\log_{10}(1)$.

As the plot shows, the majority of PIPs predictions fell just below $\log_{10}(1)$ (i.e. 0). This large number of scores at the same threshold is likely to have caused the flattening of full ROC curve in Figure 4.3 between (0.45, 0.1) and (0.5, 0.6). It is likely that as the smaller blind test sets with 1000 positives and 1000 negatives did not show this effect, increasing the sample size included more pairs with scores in this range.

In order to look more closely at the top scoring predictions, the ROC100 curves for the four predictors, shown in Figure 4.5, were plotted.

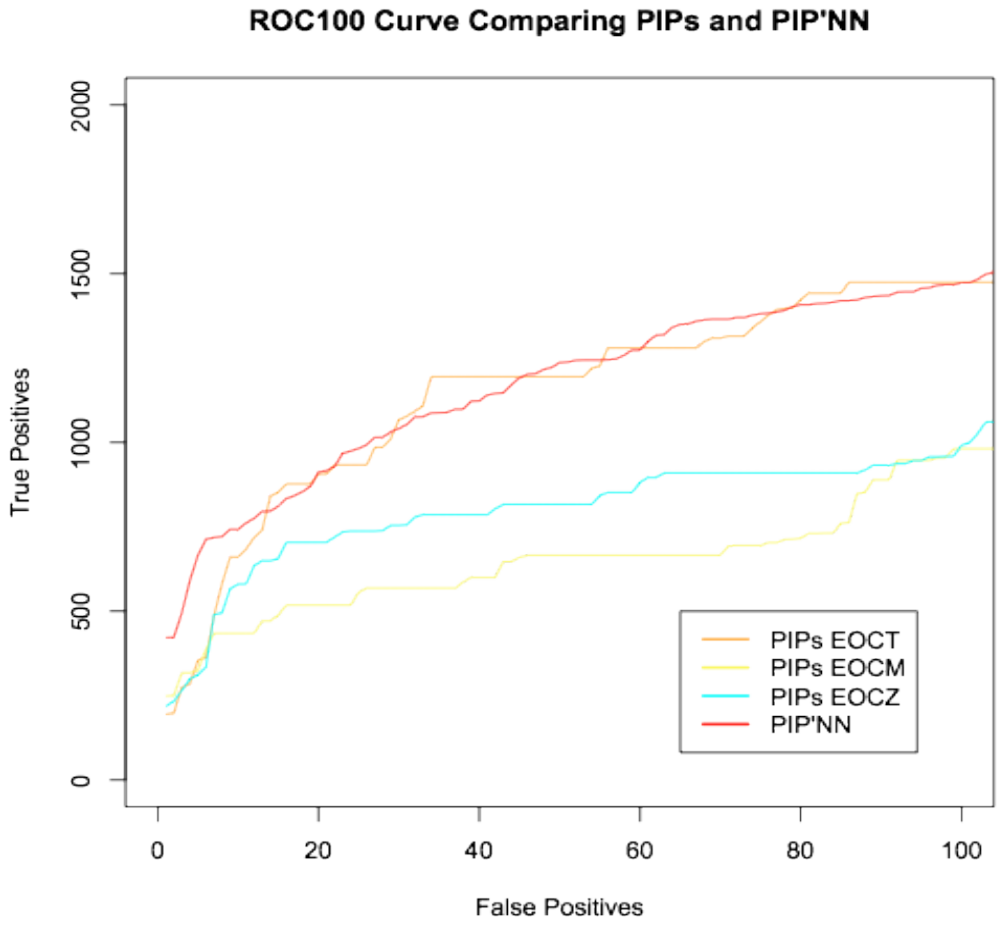


**Figure 4.5: ROC100 curve comparing PIPs and PIP'NN.** The ROC100 curves for the PIPs EOCT (orange), EOCM (yellow) and EOCZ (cyan) methods are plotted against the curve for the PIP'NN predictor (red). The EOCM and EOCZ methods both predict lower numbers of true positives than the PIP'NN and EOCT predictors. While the numbers of true positives at 100 false positives for the PIP'NN and EOCT methods are comparable, PIP'NN predicts a much higher number of true positives (796) before the first 13 false positives than the EOCT predictor (741).

While the PIPs EOCM (yellow) and EOCZ (cyan) methods predict much lower numbers of true positives before the first 100 false positives (985 and 981, respectively), the PIP'NN (red) and PIPs EOCT (orange) methods both predict comparable numbers of true positives at the 100th false positive (1473 and 1474, respectively). Crucially, however, PIP'NN scores almost double (421) the number of positives than PIPs (194) with the highest scores. Additionally, PIP'NN continued to correctly predict a larger number of true positives (796) before the first 13 false positives than the PIPs EOCT

predictor (741). This pattern suggests that the highest scoring predictions from PIP'NN are more likely to be true positives than PIPs, but as the thresholds for each predictor decrease, the accuracy of both methods becomes comparable.

Overall, these plots offer two conclusions. First, while the performance of the Bayesian PIPs predictors is highly related to the set of the proteins within the test set, PIP'NN was able to predict accurately and consistently across larger datasets. Second, although both predictors perform comparably at lower thresholds, PIP'NN is more accurate than all PIPs methods at assigning positive predictions the highest output scores.

### 4.3.2 Further Analysis of Blind Test Set Predictions

In order to analyse these blind test results further, the sets of true and false positive predictions that were unique and shared by the predictors were determined and are provided in Table 4.3.

| | True Positives | False Positives |
|---|---|---|
| **Unique PIP'NN** | 863 | 22 |
| **Unique PIPs: EOCT** | 83 | 4 |
| **Shared** | 323 | 2 |

**Table 4.3: Unique and shared true and false positive counts from the full blind test for the PIPs and PIP'NN predictors.** Counts of true and false positives predicted uniquely by the PIP'NN (row 1) and PIPs (row 2) predictors. Additionally, the numbers of overlapping true and false positive predictions are given (row 3).

While the low number of false positive predictions for the PIPs EOCT predictor is better than the number for the PIP'NN predictor (with a false discovery rate (FP/(TP+FP)) of 0.9% for PIPs EOCT and 1.8% for PIP'NN), this low rate is likely due to the low number of positive predictions (a total of 413) overall. Therefore, while PIP'NN is able to predict more positive pairs correctly, it does so with some compromise to the accuracy.

Of the 24 false positive predictions from PIP'NN, 20 had output scores in the range of 0.5 to 0.7. Of the remaining four, shown in Table 4.4, below, two scored 0.78, one scored 0.80 and two scored above 0.9.

| Prot1 | Prot2 | Spear | Pear | Ortho | Dom | GO | PTM | PIP'NN Score | PIPs Final Score |
|---|---|---|---|---|---|---|---|---|---|
| STAMBP | POLR2A | 0.620 | 0.440 | 5.54 E-04 | 0.0 | 0.0 | 0.02 | 0.779 | 0.006 |
| RELA | DDX54 | 0.0 | 0.0 | 6.52 E-04 | 7.96 E-07 | 0.0 | 0.0357 | 0.805 | 0.006 |
| LOC389000 | C20orf43 | 0.0 | 0.0 | 0.055 | 0.0 | 0.0 | 0.0 | 0.915 | 0.076 |
| HNRNPH1 | PARP1 | 0.677 | 0.532 | 5.54 E-04 | 0.0 | 0.0 | 0.085 | 0.999 | 0.008 |

**Table 4.4: Highest scoring false positive predictions in the full blind test set for the PIP'NN predictor.** Details of the four highest scoring false positives predicted by PIP'NN in the full blind test set along with the normalised values supplied to the predictor as input. Additionally, the final PIPs EOCT score (adjusted for the 1/1000 prior odds ratio) is given for each prediction in the far right column.

Looking at the input scores for each of these four pairs suggests predictions were supported by one or two moderately to highly contributing sources of evidence. Interestingly, for the prediction between LOC389000 and C20orf43, two proteins with little known evidence, the Orthology score appears to have driven the prediction. In PIPs, the Orthology module likelihood ratio for this pair (85.2), represents a pair of

proteins for which the orthologues for the each protein in the pair are known to interact in one species. As the Orthology module is typically a strong indicator of interaction, it is possible that this pair might be an interaction that has not yet been confirmed.

The scores of the remaining 20 false positives that fell in the lower scoring range closer to the prediction threshold suggest that taking 0.5 as a stringent cut-off is likely going to include a number of incorrect predictions.

### 4.3.3 Comparison of the Final Prediction Sets of the SNNS and Bayesian PIPs Predictors

The numbers of overlapping and distinct interactions predicted from the PIP'NN and PIPs EOCT, EOCM and EOCZ predictors were considered. Table 4.5 shows the number of interactions predicted by each prediction method on its own (diagonal across the table, bolded in black) and the number of interactions predicted by each combination of PIPs method and PIP'NN (far right column, bolded in red).

| | PIPs: EOCT | PIPs: EOCM | PIPs: EOCZ | PIP'NN |
|---|---|---|---|---|
| **PIPs: EOCT** | **240,560** | 128,221 | 224,917 | **117,274** |
| **PIPs: EOCM** | | **162,157** | 153,519 | **75,235** |
| **PIPs: EOCZ** | | | **579,247** | **215,235** |
| **PIP'NN** | | | | **1,999,770** |

**Table 4.5: Overlapping and total predictions in the PIPs and PIP'NN total prediction sets.** Numbers of overlapping and distinct predictions in the total prediction sets for the PIPs EOCT, EOCM and EOCZ predictors and the PIP'NN predictor. For the PIPs predictors, interactions were considered if they had a final EOCT, EOCM or EOCZ score above 1.0, after adjustment for the 1/1000 prior odds ratio. For the PIP'NN predictor, pairs were predicted as interacting with scores output scores above 0.5. The counts in the diagonal boxes through the table represent the total numbers of predictions from each predictor on its own. Each of the other

table cells then contains the number of interactions predicted by the union of the predictor in the column heading and the predictor in the column row.
Of the four predictors, PIP'NN predicts the highest number of interactions with 1,999,770 pairs with output scores above 0.5. PIP'NN predicted 54,562 of the 126,107 interaction predictions (43.3%) shared between the three PIPs predictors.

In order to see if the overlapping predictions followed a pattern of scoring, the PIPs and PIP'NN scores for each of PIPs EOCT-PIP'NN, PIPs EOCM-PIP'NN and PIPs EOCZ-PIP'NN prediction sets were compared. As shown in Figure 4.6, there is not a strong correlation across the range of PIPs and PIP'NN predictions scores within this set for each of the PIPs EOCT (Pearson's Correlation Coefficient (PCC) = 0.038), EOCM (PCC = 0.037) and EOCZ (PCC = 0.016) methods. However, the clustering of points in the upper-right corner indicates that the interactions that have the highest PIPs final scores are also assigned the highest PIP'NN scores.

| Protein1 | Protein2 | EOCT Score | EOCM Score | EOCZ Score | PIP'NN Score |
|---|---|---|---|---|---|
| PSMA2 | PSMB4 | 553420 | 457777 | 275141 | 1.0 |
| SNRPE | SNRPF | 482770 | 231131 | 240017 | 1.0 |
| JUN | JUNB | 363670 | 300262 | 194548 | 1.0 |
| SNRPD2 | SNRPF | 280511 | 134297 | 139460 | 1.0 |
| CFL1 | ACTG1 | 241264 | 167631 | 129066 | 1.0 |
| CDC2 | CDC25C | 214595 | 199198 | 148603 | 1.0 |
| PSMB3 | PSMA2 | 199360 | 164907 | 99115.2 | 1.0 |
| PSMA2 | PSMA4 | 199360 | 164907 | 99115.2 | 1.0 |
| STAT5A | STAT5B | 141998 | 22970.3 | 18663.8 | 1.0 |
| JUN | JUND | 121959 | 61631.5 | 39932.8 | 1.0 |

**Table 4.6: Prediction scores for the top ten overlapping predictions from the three PIPs and PIP'NN predictor.** The table above is an extension of the table in Chapter 2.3.4: Top Predictions for the ten highest scoring predictions from the EOCT, EOCM and EOCZ PIPs predictors. The output score from the PIP'NN predictor is given in the far right column for each pair in the set.

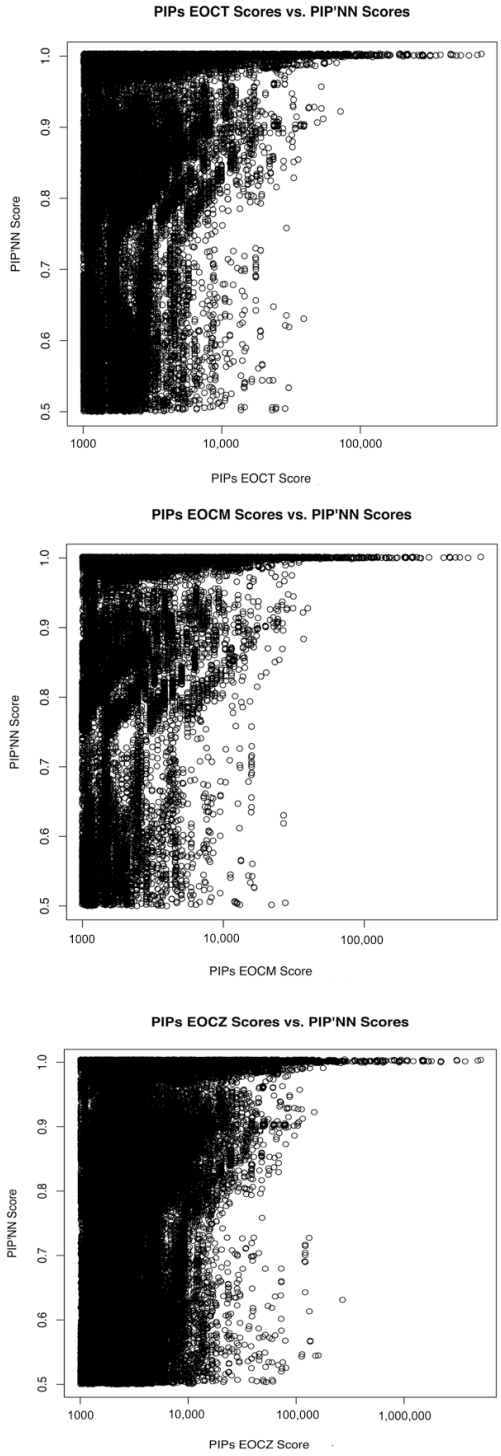


**Figure 4.6: Correlation between PIPs EOCT, EOCM and EOCZ final scores and PIP'NN output scores.** The scatterplots showing the relationship between PIPs EOCT (A, $PCC_{EOCT}$ = 0.038, df=117272, t=13.028, p-value=2.2e-16), EOCM (B, $PCC_{EOCM}$ = 0.037, df =75233, t=10.0286, p-value=2.2e-16) and EOCZ (C, $PCC_{EOCZ}$ = 0.016, df=215233, t=7,4245, p-value=1.14e-13) likelihood ratios and PIP'NN output scores are provided.

This observation is confirmed by the top ten overlapping predictions, shown in Table 4.6, above, with their scores for the three PIPs methods along with their output score from PIP'NN. While not shown, the remaining 40 predictions were also predicted as interacting with PIP'NN with scores of either 1.0 or a slightly lower 0.999, suggesting that for the highest scoring predictions from PIPs, there is an agreement of predictions with the PIP'NN predictor.

Figure 4.7 plots the distributions of PIP'NN scores for each set of overlapping PIPs EOCT-PIP'NN, PIPs EOCM-PIP'NN and PIPs EOCZ-PIP'NN.

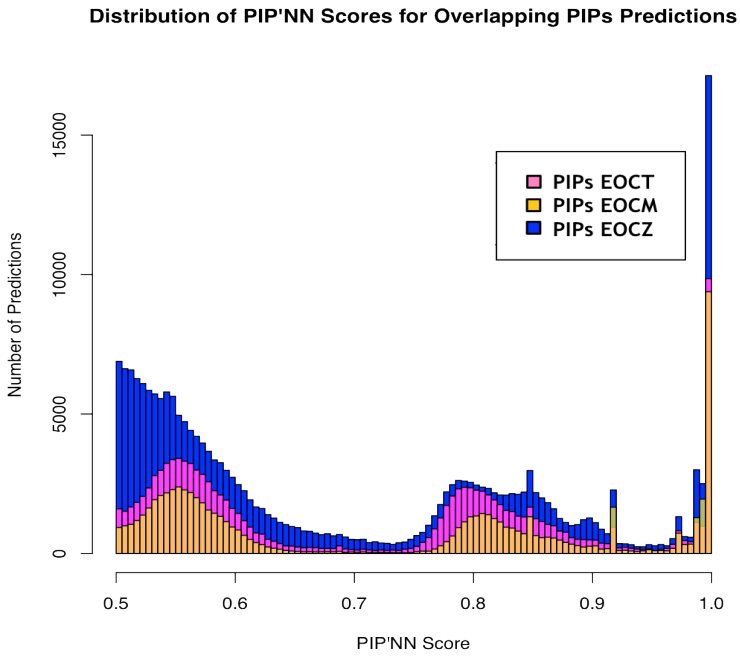


**Figure 4.7: Histogram distribution of PIP'NN scores for predictions overlapping with the PIPs EOCT, EOCM and EOCZ prediction sets.** Pairs with PIP'NN output scores above 0.5 and PIPs EOCT, EOCM or EOCZ scores above 1.0 were selected, and the distribution of PIP'NN scores for each set were plotted (PIPs EOCT-PIP'NN: pink, PIPs EOCM-PIP'NN: yellow, PIPs EOCZ-PIP'NN: blue). All three distributions follow a similar pattern of most interactions predicted in either a 'low' range (between 0.5 and 0.65), a 'mid' range (between 0.75 and 0.9) or a 'high' range (1.0).

Interestingly, the PIP'NN scores for these sets of predictions appear to be clustered in three ranges: a 'low' range (between 0.5 and 0.65), a 'mid' range (between 0.75 and 0.9) and a 'high' range (1.0). Table 4.7 shows the number of PIPs predictions that fall in each of these ranges.

| PIPs Method | Low PIP'NN (0.5-0.75) | Mid PIP'NN (0.75-0.9) | High PIP'NN (0.9-1.0) |
|---|---|---|---|
| **PIPs EOCT** | 59,736 | 39,844 | 15,080 |
| **PIPs EOCM** | 37,188 | 20,675 | 15,083 |
| **PIPs EOCZ** | 126,354 | 55,517 | 30,198 |

**Table 4.7: Number of PIP'NN predictions in low, mid and high score ranges.** The number of predictions with PIP'NN scores in a low range (0.5-0.75), mid range (0.75-0.9) and high range (0.9-1.0) with PIPs EOCT, EOCM or EOCZ final scores above 1.0 are provided.

Considered together, Figures 4.5 and 4.6 and Table 4.7 offer a suggestion for how both the PIPs and PIP'NN output scores could be considered together to add an extra filter to predictions from either method on its own. For example, if predictions from PIP'NN with low output scores (i.e. between 0.5 and 0.75) also had a PIPs EOCT, EOCM or EOCZ prediction score above 1.0, they could be flagged as a 'top hit'. Likewise, predictions from one of the PIPs methods with low final scores (i.e. those just above 1.0) could also be flagged as a 'top hit' if they have PIP'NN scores above 0.5.

### 4.3.4 Performance of PIPs and PIP'NN on Known Negative Interactions

Without a full source of known, negative interactions, it is difficult to construct a blind test set with full certainty that the examples included in the negative dataset are truly non-interacting. While one resource does exist, the Negatome (Smialowski *et al.*, 2010)

(see Chapter 1.5.1.2: Negative Datasets), it currently only contains 1291 protein pairs. To test how well PIPs and PIP'NN were able to correctly predict these pairs as non-interacting, the EOCT, EOCM and EOCZ final likelihood ratios and PIP'NN output score were obtained and are compared in Table 4.8.

| Method | Number of False Positives |
|---|---|
| PIPs EOCT | 65 |
| PIPs EOCM | 71 |
| PIPs EOCZ | 110 |
| PIP'NN | 183 |

**Table 4.8: Number of pairs in the Negatome incorrectly predicted as interacting.** Pairs were considered interacting for the PIPs methods if they had final likelihood ratios (before adjustment for the prior odds ratio) above 1000.0 or PIP'NN output scores above 0.5.

While the PIPs EOCT and EOCM methods perform comparably, the PIPs EOCZ method and PIP'NN predict approximately double and triple the number of interactions. Of the PIP'NN predictions, 63 were also predicted as interacting by one of the PIPs methods; of these 63 predictions, all by seven had PIP'NN output scores of 0.99 or 1.0. Of the remaining 130 predictions, 36 also had PIP'NN scores of 0.9 or higher. That the majority of predictions from PIPs were also assigned the highest PIP'NN scores suggests that these predictions were based on evidence that was strong enough across the Expression, Orthology and Combined modules to be considered by both methods. For the remaining PIP'NN predictions, particularly those with the highest scores, the lack of a matching PIPs prediction likely indicates that despite evidence suggesting an interaction, it was either not strong enough or was assigned a low likelihood ratio in one

of the Transitive, Cluster or TransMCL modules. It could also be taken into consideration that the interactions recorded in the Negatome were selected, at some point, for validation with lab experimentation; therefore, it is likely that enough evidence for possible interaction was present to warrant further investigation.

### 4.3.5 Comparison of PIPs with Other Predictors of Human Protein-Protein Interaction

Currently, there are several other predictors of human protein-protein interaction prediction available for public use. While each method each capitalises on a range of similar comparative genomics and gene conservation principles with the common goal of identifying potentially novel interactions, each also differs in its training and testing dataset construction, evidence considered, method of analysis and final results scoring method. Due to these differences, benchmarking all methods against one another in a fully comprehensive and unbiased way is a difficult task. However, an understanding of how PIPs and PIP'NN perform in comparison in line with the current field is an important assessment of its usefulness in practical application.

A brief description of the five methods selected for comparison (STRING (Szklarczyk *et al.*, 2011), PrePPI (Zhang *et al.*, 2012), FunCoup (Alexeyenko *et al.*, 2012), IntNetDBv.1.0 (Xia, Dong & Han, 2006) and BIPS (Garcia-Garcia *et al.*, 2012)) is provided in Table 4.9.

| PPI Method | Species | Brief Description | Evidence Considered | Scoring Method | Prediction Cut-Off |
|---|---|---|---|---|---|
| **STRING** (http://string-db.org/) | 1133 different organisms | Includes known and predicted protein interactions. | All interactions are assigned a confidence score according to the KEGG database based on whether the pairs are likely to be in the same pathway.<br>Includes evidence on conserved gene neighbourhood, gene fusion events and gene co-occurrence across genomes.<br>Two methods of orthology transfer: COG method or protein-mode.<br>Large-scale experiments. | Naïve Bayesian Network<br>Each link between two proteins in the network is assigned a confidence score. | Low confidence: score < 0.4<br>Mid confidence: score >= 0.4 and score < 0.7<br>High confidence: score > 0.7 |
| **IntNetDB v1.0** (http://hanlab.genetics.ac.cn/sys/intnetdb/) | *H. sapiens, C. elegans, M. musculus, D. melanogaster* | The probability of an interaction occurring between two proteins computed through a naïve Bayesian network. | Gene co-expression, binary protein-protein interactions between orthologues, shared neighbours, phenotype similarity, shared GO annotations, domain-domain interactions and gene context | Naïve Bayesian Network<br>Prediction scores presented as $\log_2$LR (LLR) values. | LLR = 6.0 (44.1% confidence)<br>LLR = 7.0 (55.6% confidence)<br>LLR = 9.0 (81.1% confidence)<br>LLR = 16.0 (97.8% confidence) |
| **FunCoup** (http://funcoup.sbc.su.se) | *H. sapiens*, *S. cerevisiae*, *D. melanogaster*, *R. norvegicus*, *M. musculus*, *C. elegans* | Transfers information from model organisms (*M. musculus*, *C. elegans*, *S. cerevisiae*, *D. melanogaster*, etc.) via InParanoid orthologues to identify functional coupling between proteins | Whole protein and domain contacts, mRNA and protein expression, localisation in tissues and cellular compartments, miRNA and transcription factor targeting, similar phylogenetic profiles | Naïve Bayesian Network<br>Scoring:<br>where FBS > 3 is considered predicted. FBS also corresponds to a confidence score (PFC). | Final score = 6.9 (PFC = 0.5)<br>Final score = 7.9 (PFC = 0.75)<br>Final score = 9.9 (PFC = 0.95) |

| | | | | | |
|---|---|---|---|---|---|
| **BIPS** (http://sbi.imim.es/web/index.php/research/servers/bips) | *H. sapiens* | Interolog principle: If proteins A and B interact in one species, then their orthologues A' and B' are likely to interact in another species.<br><br>Also includes a second approach where query proteins are compared to PFAM domain sequences, and if the domains interact, the proteins are likely to interact. | Sequence and domain homology | Orthologues and PFAM domain mapping of individual proteins are determined by BLAST e-values. | Only predicted interologs included in prediction set. |
| **PrePPI** (http://bhapp.c2b2.columbia.edu/PrePPI/) | *H. sapiens* | Aligns sequences of query proteins with sequences of proteins with structural models or homology domains and scores according to similarity of query-template. | Sequence-based structural alignments<br><br>Essentiality of proteins in the interacting pair, co-expression level, GO functional similarity, Munich Information Centre for Protein Sequences (MIPS) functional similarity and phylogenetic profile similarity. | Naïve Bayesian Network | Likelihood ratio > 600.0 |

**Table 4.9: Brief details of the five methods compared.** Details are provided for the range of species covered in the method, a brief description of the background theory, sources of evidence considered for predictions, the learning method implemented and the current web server address (Brown & Jurisica, 2005; 2007; Szklarczyk *et al.*, 2011; Alexeyenko *et al.*, 2012; Garcia-Garcia *et al.*, 2012) (Zhang *et al.*, 2012).

Consideration of this comparison requires acknowledgement of several points. First, a true comparison of the performance of each of the methods above is impossible without complete access to training datasets to ensure that the test is truly blind. Therefore, it is possible that some of the methods have been trained on pairs in the comparison set chosen. Likewise, for prediction methods like STRING, which double as a database of all known protein interactions, a true assessment of how well these interactions would have been identified had they not already been included in the interaction set is not possible. Second, due to the differences in prediction algorithms, sources of data and protein pairs considered by each method, 100% coverage of the selected dataset across the five methods was impossible. While care was taken to map the full set of proteins considered by each method to their UniprotAC identifiers, it was not possible to match every one. Therefore, it is possible that some interactions have been excluded from each method that were included in the comparison set chosen.

Like PIPs and PIP'NN, STRING, FunCoup, IntNetDB and BIPS all incorporate either gene context or orthology-based evidence without consideration of sequence or structure. As these methods are based on similar sources of evidence, they were first compared against each other. Table 4.10 provides the number of pairs included in and able to be matched from each available resource and how many of these pairs were predicted as interacting according to the criteria for each method. Figure 4.8 shows a barplot comparing the number of pairs preidcted as interacting by each of the methods considered.

| PPI Method | Proteins Considered | Predictions |
|---|---|---|
| PIPs | 748 | At LR cut-off = 1000.0:<br>71 total<br>51 from EOCT, 38 from EOCM and 64 from EOCZ with 33 predicted by all.<br>At LR cut-off = 400.0:<br>157 total<br>72 from EOCT, 58 from EOCM, 103 from EOCZ with 38 predicted by all. |
| PIP'NN | 748 | At final score > 0.5: 102 total<br>At final score > 0.7: 65 total |
| STRING | 748 | 231 including HPRD, 77 excluding HPRD.<br>Of the 77:<br>At score < 0.4 (low confidence): 25 total<br>At score >= 0.4 and < 0.7 (mid confidence): 8 total<br>At score >= 0.7 (high confidence): 44 total |
| IntNetDB v. 1.0 | 84 | At LR cut-off = 6.0 (55.6% confidence): 84 total<br>At LR cut-off = 9.0 (81.1% confidence): 40 total |
| FunCoup | 89 | At FBS score > 6.9 (PFC of 0.5): 62 total<br>At FBS score > 7.9 (PFC of 0.75): 55 total<br>At FBS score > 9.9 (PFC of 0.95): 42 total |
| BIPS | 1 | 1 interaction included in available dataset |

**Table 4.10: Comparison of number of protein pairs considered and number of predicted interactions between PIPs and PIP'NN and four other predictors.** Column two shows the number of proteins in the selected test dataset (748 pairs total) able to be matched in the PIPs, PIP'NN, STRING, IntNetDB v.1.0, FunCoup and BIPS databases. Column three gives the number of pairs out of the number in column two that score above the prediction cut-off thresholds for each method.

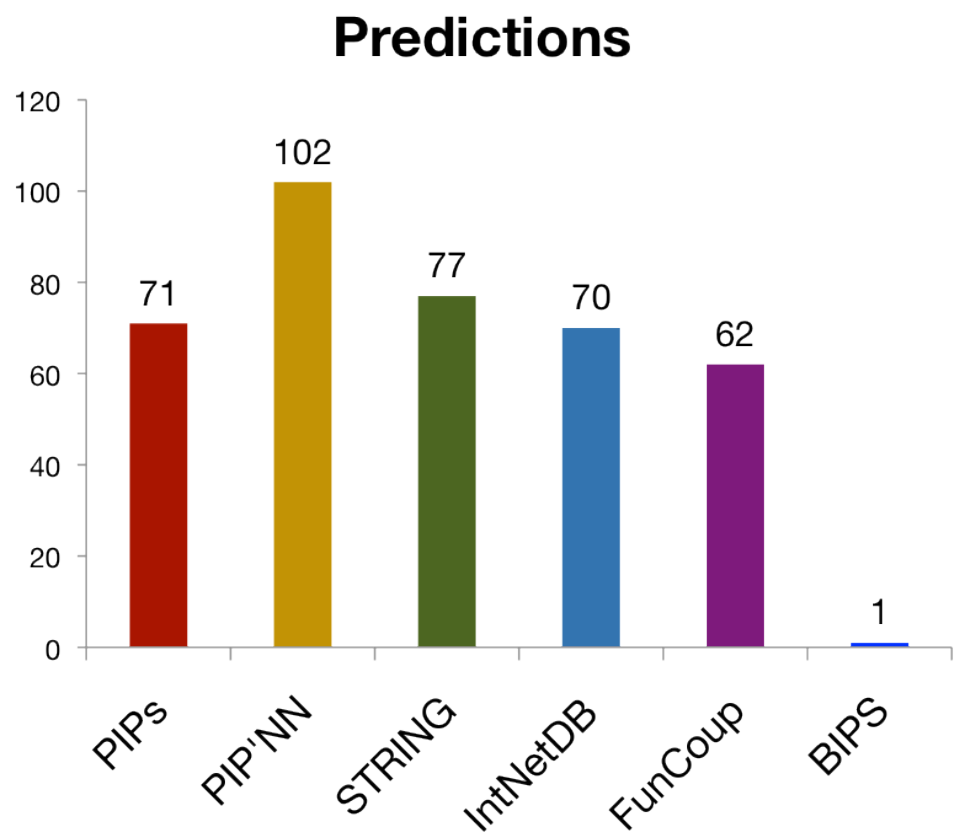


**Figure 4.8: Barplot comparing the number of pairs predicted as interacting by each of the six methods considered.** Pairs were considered 'predicted as interacting' if they fell above the specific threshold for each method (see Tables 4.1 and 4.10, above).

As expected, the differences between the learning method and output reporting made it difficult to directly compare results. First, the low number of predictions for BIPS, which bases its algorithm mainly on the orthologous transfer of interactions, is most likely due to the limited data coverage allowed by only examining protein pairs in other species known to interact. For FunCoup, pairs were only included in the downloadable prediction dataset if they had a final score above the significance threshold of 4.0. Of the 89 interactions in the comparison test set that met this criteria, 42 (47%) had final FBS scores above 6.9, which corresponds to a 95% confidence that the interaction occurs. IntNetDB, which employs a naïve Bayesian network similar to the PIPs framework, scored 84 of the interactions in the comparison set with $\log_2$ likelihood

ratios (LLR) above 6.0 (56% confidence), of which 42 had scores above 9.0 (81% confidence).

As STRING serves as a database of both known interactions and interactions predicted based on evidence of gene conservation and co-occurrence, assessing the method from an interaction prediction standpoint proved slightly more difficult. Of the 748 interactions in the test set, 231 were able to be mapped to interactions in STRING and were recorded as 'binding' (i.e. indicating a physical interaction and not just a functional relationship). However, of this set, 154 were annotated as having been included in the database based on the HPRD and were therefore removed. Of the remaining 77 interactions, all but one were annotated as having been included based on either the Reactome, MINT, KEGG, PID or PDB databases, allowing 44 to be assigned high confidence scores, 8 mid confidence scores and 25 low confidence scores.

| | PIP'NN | PIPs | STRING | FunCoup | BIPs | IntNetDB |
|---|---|---|---|---|---|---|
| **PIP'NN** | **102** | **48** | **42** | **27** | **0** | **46** |
| **PIPs** | | **71** | **20** | **27** | **1** | **41** |
| **STRING** | | | **77** | 6 | 0 | 42 |
| **FunCoup** | | | | **62** | 1 | 27 |
| **BIPs** | | | | | **1** | 0 |
| **IntNetDB** | | | | | | **84** |

**Table 4.11: Overlap of predictions by individual methods.** The number of predictions, out of the 748 in the prediction comparison set, made by each method are shown diagonally in bolded black. The number of these predictions that overlapped with the predictions made by PIP'NN are shown in bolded red, and the number of predictions that overlapped with predictions made by PIPs are shown in bolded blue. The number of these predictions that overlapped between the other methods are shown in normal black.

Table 4.11, above, shows the number of predictions overlapping between pairs of methods. Among the other methods considered, PIPs shared the highest number of overlapping predictions with PIP'NN (48), IntNetDB (41), FunCoup (27) and STRING (20), while PIP'NN shared the highest number of overlapping predictions with PIPs (48), IntNetB (46) and FunCoup (27). PIP'NN also identified the highest number of overlapping interactions with each of the other four methods.

Overall, PIPs and PIP'NN predict similar numbers of known interactions within the test dataset to STRING, IntNetDB v. 1.0, FunCoup and BIPS. Additionally, these interactions overlap with the interactions predicted by the other methods. This analysis suggests that PIPs and PIP'NN perform comparably to other predictors of human protein-protein interactions that consider similar sources of evidence.

Recently, a new method, PrePPI (Zhang *et al.*, 2012), has been developed. Like PIPs, PrePPI incorporates a similar Bayesian network framework and similar sources of gene and orthology evidence. However, unlike PIPs, PIP'NN and the other four methods compared above, PrePPI also considers structure as an additional source of evidence. While previous attempts have been made to include structure into protein-protein interaction prediction methods (Bock & Gough, 2001), these methods have been limited by the number of known structures. As an alternate approach, PrePPI uses sequence alignment to assign an either actual structure or homology model to each protein. The interaction models are then scored with five different criteria based on how similar the protein models are to their template proteins and on the conservation and properties of the interaction interface (Zhang *et al.*, 2012). The final score from this structure

analysis forms one component of the naïve Bayesian network to determine the final likelihood ratio.

When predictions were obtained from PrePPI for the comparison dataset, the method far outperformed all other methods considered, including PIPs and PIP'NN, with a total of 614 pairs predicted with final likelihood ratios above their selected cut-off of 600.0 and 562 pairs above the PIPs cut-off of 1000.0. As the other sources of evidence considered by PrePPI (co-expression, GO functional similarity and phylogenetic profile similarity) overlap with those considered by PIPs, it is clear that the inclusion of structure has contributed the prediction of such a higher number of interactions. Based on these results alone, this method of structural modelling suggests a new direction for both PIPs and the entire field to allow the inclusion of structure into protein-protein interaction prediction frameworks.

However, while identifying known positives is key, it is also necessary to consider the rate of false positive predictions for a predictor. Therefore, as an additional comparison, the number of the interactions included in the Negatome (see Section 4.3.4, above) predicted by each of the methods above were also obtained and are provided in Table 4.12, below.

Of the methods considered, PrePPI predicted the largest number of pairs in the Negatome dataset as interacting (218) at likelihood ratios above its selected cut-off of 600.0, followed by PIP'NN at a cut-off threshold of 0.5 (183) and IntNetDB v.1.0 at a cut-off threshold of 6.0 (145). Therefore, in order to further investigate the accuracy of PrePPI compared to PIPs and PIP'NN, a full ROC (Figure 4.9) curve was plotted by

taking the original comparison set as a positive dataset and the Negatome set as a negative dataset.

| Method | Number of Predictions Included in the Negatome |
|---|---|
| PIP'NN | At Output Score >= 0.5: 183 total<br>At Output Score >= 0.7: 104 total |
| PIPs | At LR=1000.0:<br>EOCT: 56 total, EOCM: 62 total, EOCZ: 104 total |
| STRING | At score < 0.4 (low confidence): 25 total<br>At score >= 0.4 and < 0.7 (mid confidence): 8 total<br>At score >= 0.7 (high confidence): 44 total |
| FunCoup | No Interactions Matched |
| BIPs | No interactions |
| IntNetDB v.1.0 | At LR cut-off = 6.0 (55.6% confidence): 145 total<br>At LR cut-off = 9.0 (81.1% confidence): 99 total |
| PrePPI | At LR=600.0: 218 total<br>At LR=1000.0: 204 total |

**Table 4.12: Number of Predicted Interactions included in the Negatome.** The number of interactions predicted by PIP'NN, PIPs, STRING, FunCoup, BIPs, IntNetDB v.1.0 and PrePPI that are included in the Negatome Database (Smialowski *et al*., 2010) were obtained at the stated thresholds.

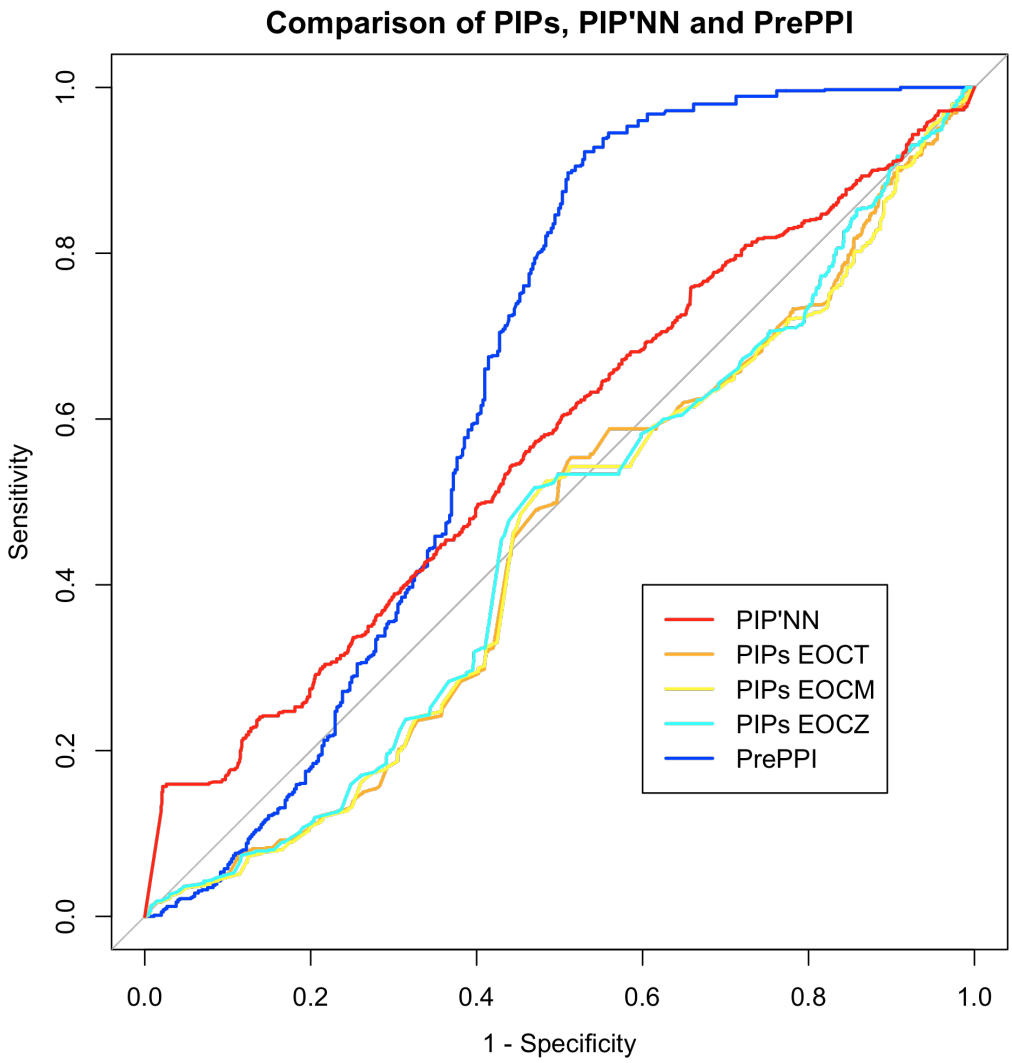


**Figure 4.9: Full ROC curve comparing the performance of PIPs, PIP'NN and Preppi on the positive and negative datasets.** The full ROC curves for PIP'NN (red), the PIPs EOCT (orange), EOCM (yellow) and EOCZ (cyan) methods and PrePPI (blue) are plotted. The positive dataset included 748 new interactions added to the HPRD between August 2010 and August 2011 (described above), and the negative dataset included 1211 interactions included within the Negatome database (Smialowski *et al.*, 2010). While the three PIPs methods (EOCT AUC=0.455, EOCM AUC=0.451 and EOCZ EUC=0.461) perform worse than random (grey line), PIP'NN (AUC=0.571) and PrePPI (AUC=0.651) perform marginally better on the test.

To get a clearer picture of the performance of PIPs, PIP'NN and PrePPI at the highest scoring interactions, the ROC50 curves (Figure 4.10) for the five predictors were also plotted.

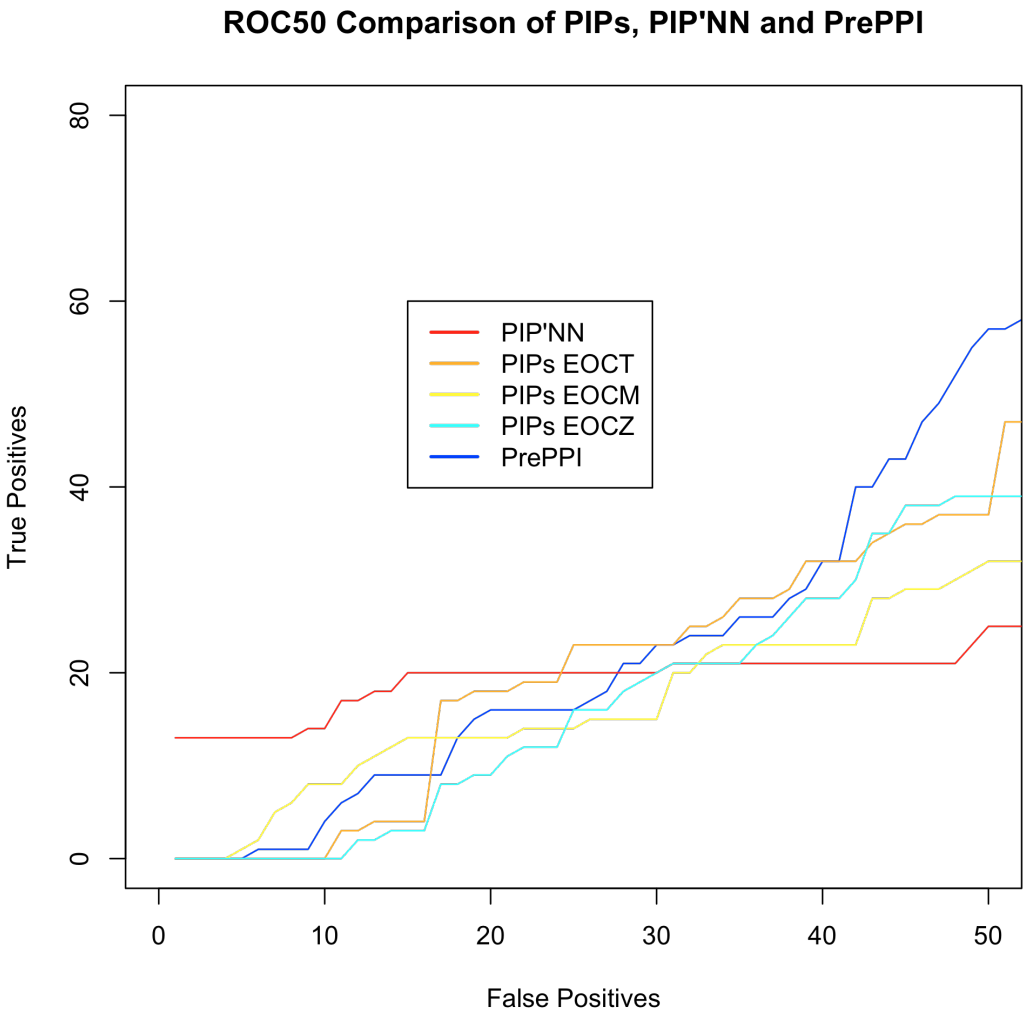


**Figure 4.10: ROC50 curves comparing the performance of PIPs, PIP'NN and Preppi on the highest scoring positive and negative predictions.** The ROC50 curves for the three PIPs methods (EOCT-orange, EOCM-yellow and EOCZ-cyan), PIP'NN (red) and Preppi (blue) were plotted. While Preppi ultimately predicts the greatest number of known positives before the 50$^{th}$ false positive (57), PIP'NN predicts a greater number of positives with higher scores than any of the pairs included in the Negatome (16).

While PrePPI predicts a greater number of interactions within the positive dataset at scores higher than the 50$^{th}$ pair in the negative dataset, the highest scoring predictions of both datasets are from the Negatome. Although PIP'NN ultimately predicts a lower number of pairs within the positive dataset as interacting than PrePPI, it assigns a

greater number of pairs within the positive dataset (13) before the first negative. Looking further at the distribution of scores assigned to the Negatome predictions from the two methods (not shown), PIP'NN assigns 31% (58 of 183) of pairs scores above a stringent threshold of 0.8, while PrePPI scores 76% (165 of 218) with high likelihood ratios above 10000.0. While the number of Negatome predictions from PIP'NN decreases considerably as the threshold increases, the majority of predictions from PrePPI score are predicted with very high likelihoods. Overall, these score distributions, along with the PIP'NN and PrePPI ROC50 curves, suggest that although PrePPI predicts a greater number of positives overall, PIP'NN is more accurate with assigning the highest scores to known positive interactions.

## 4.4 Discussion

Overall, assessment of the accuracy and total prediction sets for the PIPs and PIP'NN predictors offer several conclusions. First, while little difference was observed initially between the prediction accuracy of the PIPs and PIP'NN predictors on the small, original blind test set, increasing the size of the dataset altered results and gave a more accurate assessment of how each of the predictors performed across a larger set of examples. Strictly in terms of performance, PIP'NN is able to score known positive and negative interaction examples in a large blind test set more accurately and consistently than the EOCM and EOCZ PIPs prediction methods. Compared the PIPs EOCT method, PIP'NN is able to assign a greater number of true positive examples in a large blind test with very high output scores, but its accuracy becomes more comparable at lower cut-off thresholds.

While there is an overlap of over 28% between the PIPs EOCT and PIP'NN final prediction sets at a cut-off threshold of 0.6, the number of predicted interactions from PIP'NN is over three times higher. With this higher coverage comes the risk of large numbers of false positive predictions; however, when assessing the negative examples predicted as interacting in the blind test set, the number that PIP'NN predicted is still relatively low. Furthermore, of these false positive predictions, the majority had final output scores that hovered at or around the selected prediction cut-off threshold.

Examining the sets of overlapping predictions from the three PIPs methods (both individually and together) and PIP'NN has offered a suggestion for how PIPs and PIP'NN could be employed together in a prediction framework. While the plot of PIPs versus PIP'NN scores (Figure 4.5) shows that there is concordance between both methods for the highest scoring predictions, there is less of a correlation between the lower scoring predictions. For these interactions, particularly those slightly above the PIPs or PIP'NN score cut-off thresholds, having two methods of analysis will help filter the prediction set down to the most likely interactions. For example, the large number of interactions predicted by PIP'NN that fall in the mid-range of 0.4-0.6 could be assessed further by their PIPs score to add an extra layer of confidence to predictions. However, the unique interactions predicted by each method independently should not be discounted immediately before they are assessed practically for viability. Overall, this combination of predictors will allow for both prediction sets to be filtered down to the most likely interactors and for pairs that might have been missed by either method due to the differences in how the available evidence for the pair is handled.

Finally, comparing PIPs and PIP'NN to five other current protein-protein interaction prediction methods highlighted two main conclusions. First, both PIPs and PIP'NN performed comparably to the four methods that considered similar sources of evidence and did not include structure (i.e. STRING, IntNetDB, FunCoup and BIPS) and predicted similar interactions. Second, the fifth method considered, PrePPI, which does consider structure in addition to the other features covered by PIPs, predicted a substantially higher number of interactions than all of the other methods.

However, that PrePPI predicts the greatest number of pairs within the Negatome as interacting suggests that this high rate of positive predictions comes at some cost to its accuracy. In particular, the distribution of scores predicted by PIP'NN, in which the majority of Negatome predictions fall in the lower range of PIP'NN output scores, compared to the distribution of scores predicted by PrePPI, in which most predictions are scored with very high likelihood ratios, suggests that PIP'NN is better able to not predict false positives with the highest scores.

While this comparison of existing protein-protein interaction prediction methods has shown PIP'NN in particular to be a competitive predictor among methods incorporating similar sources of evidence into their prediction frameworks, the sources for bias in this study do need to be acknowledged. Most crucially, without knowing the composition of the positive and negative datasets used for training each method, it is possible that pairs in the selected HPRD and Negatome sets are not truly blind. While the Negatome does allow a true negative dataset to be constructed, it is possible that these interactions, which likely include pairs thought to interact based on available evidence enough where they were studied structurally, have evidence supporting interaction and are thus not

representative of 'strongly' negative interactions (i.e. those with no or little evidence supporting interaction). Therefore, while this comparison does offer insight into how well these methods perform against each other, it cannot be taken as absolute.

Regardless, the drastic increase in number of positives able to be predicted by including structure into PrePPI's framework has highlighted a new direction both for PIPs and PIP'NN individually and the protein-protein interaction prediction field as a whole to take in the future.

## 4.5 Conclusions

1) While on small blind test datasets there appears to be little difference between prediction accuracy of the PIPs EOCT, EOCM and EOCZ and PIP'NN methods, on a larger dataset, PIP'NN is able to correctly classify positive and negative examples more consistently than each PIPs method.

2) Although the overlap of final prediction sets between the three PIPs methods and PIP'NN is low, the overlap between just the PIPs EOCT and PIP'NN predictors is much higher at 28%. PIP'NN provides a much greater coverage of the potential protein interaction network; however, this increased coverage might also include higher numbers of false positive predictions.

3) Of the highest scoring interactions predicted by the PIPs predictors, all are also predicted by PIP'NN with the highest possible scores, suggesting that there is congruence between both methods for the most likely interactions.

4) The best way of incorporating PIPs and PIP'NN is proposed as a combination method where the PIPs EOCT predictor and PIP'NN prediction outcomes are considered with alongside, and not independently, of each other.

5) Compared to other currently available predictors of human protein-protein interaction, both PIPs and PIP'NN performed comparably to the four that do not incorporate sequence analysis into their prediction frameworks. All predictors were out-performed by PrePPI, a method similar to PIPs that does include structural modelling as an additional module. However, the number of pairs within the Negatome predicted by each of the methods showed that the high positive prediction rate of PrePPI comes at some compromise to its accuracy. Overall, these results have placed PIP'NN in particular as a top contender among other current methods of protein-protein interaction prediction and have suggested a new potential direction for both PIPs and the entire protein interaction prediction field to take in the future.

# Chapter 5

## Practical Application of the PIPs and PIP'NN Predictors

## Preface

---

This chapter describes two examples of the practical implementation of the PIPs and PIP'NN predictors in lab experimental pipelines. In the first study, the use of PIPs to predict potential interactions for a set of DNA repair proteins is described. In the second analysis, PIPs and PIP'NN are incorporated as an additional level of filtering the resulting dataset from a co-immunoprecipiation/SILAC experiment with the ubiquitin ligase CUL4B.

# 5.1 Introduction

While both the Bayesian PIPs and new neural network PIP'NN protein interaction predictors have proven able to correctly predict known positive and negative interactions, the true test of effectiveness is how well both can perform in real world application. Within the University of Dundee, there are multiple groups attempting to identify novel interactions involving specific proteins of interest who would benefit from incorporating an *in silico* method of interaction prediction into their research protocols.

Of particular interest to several groups in the College of Life Sciences are the nuclear and subnuclear cell compartments and their associated molecular processes. While vast numbers of proteins and interactions within the nucleus and nucleolus have been identified, the key processes of DNA replication and repair are far from solved. With a wealth of potential nuclear and nucleolar protein-protein pairings possible, systematically testing each with lab-based experiments alone is a costly, inefficient and near impossible task. Although eventual experimental validation of interactions is necessary, machine-learning predictive techniques serve as a method to consolidate the set of interactors for testing to a subset of the most probable interactions.

Collaborations between our group and other lab-based groups both in and out of the University are mutually beneficial. With a smaller, hand-picked set of potential interactors to test, lab-based groups can save the time and money wasted with testing large sets of unlikely interactions. Likewise, experimental confirmation of interactions

predicted by the PIPs predictors supports the validity and credibility of the PIPs framework.

We have collaborated with two groups with in College of Life Sciences to practically implement PIPs to identify potential interactions among nuclear and nucleolar proteins. In the first collaboration, we teamed up with the John Rouse group to predict interactions among a selected set of proteins involved in the human DNA homologous repair system that were then tested in the lab with co-immunoprecipitation experiments. In the second collaboration, we joined with the Angus Lamond group to incorporate predictions from PIPs into their existing framework for nucleolar protein interaction detection involving co-immunoprecipitation and SILAC mass spectrometry analysis. In particular, the merging of either or both of the PIPs predictors with the experimental, proteomics and bioinformatics techniques already established for SILAC protein interaction identification show promise for developing an even stronger method of interaction validation.

### 5.1.1 The DNA Homologous Repair System

DNA damage inducing the formation of secondary nucleotide structures or nicks in one or both of the strands can have the catastrophic effect of impeding progress of DNA replication. As a response, cells are equipped with an extensive DNA repair system that is largely conserved across eukaryotes, prokaryotes and archaea. While recent discoveries have slowly begun to uncover to the intricacies of this process of homologous DNA repair in humans, many details are still unknown and required before

a clear and accurate understanding can be fully reached (Rouse, 2009; Duro *et al.*, 2010; MacKay *et al.*, 2010).

The halting of DNA replication is instigated by a collapse of the replication machinery when either a nick in one of the DNA strands or a lesion in the parent strand is encountered, or the two strands have covalently annealed together into an inseparable DNA double-helix (interstrand crosslinking, or ICL) that blocks further movement of the replisome (Rouse, 2009). Additionally, the DNA replisome can backtrack on itself and re-anneal the parent and template strands together, causing the newly formed strands to anneal together and the formation of a 'chicken foot' structure in which the DNA contains two branches (Rouse, 2009). In all cases, double-strand breaks between the DNA occur that leave exposed single-stranded DNA ends from which replication can be resumed by the process of homologous DNA repair. Homologous repair occurs either through double-stranded DNA repair, synthesis-dependent strand annealing or formation of the double Holliday junction, all of which involve an exposed 3'-end single-strand that is exposed to a complex of repair proteins that search for a homologous sequence to act as a primer for continued DNA synthesis (Rouse, 2009).

A brief overview of the current understanding of the homologous recombination process at replication forks due to interstrand crosslinks (ICL) is shown graphically in Figure 5.1, below. After recognition of a stalled replisome and a collapsed replication fork (A and B), MUS81-EME1, a structure-specific endonuclease member of well-characterised XPF nuclease family, nicks the template strand (in green) to create a single-stranded DNA overhang at the 3'-end (C). This 3'-overhang is then coated with Replication Protein A (RPA) (D), which is displaced by RPA (E), at which point the recently

identified TONSL scaffold and its complexed MMS22L, the histones H3 and H4 and their associated ASF1A/B proteins and the MCM replicative helicase also associate with the DNA (Duro *et al.*, 2010; Cybulski & Howlett, 2011). During the next stage, the 'unhooking' of the ICL (F), the ICL is unwound, its 3'- and 5'-ends cleaved, and the 3'-overhang of the template strand is extended through DNA synthesis to fill the gap caused by the removed ICL (G and H). In the final stage, the replication fork is regenerated (I) (Rouse, 2009; Cybulski & Howlett, 2011).

The individual steps of the repair pathway have recently been shown to be mediated by two main complexes. SLX4, the protein product of one of six genes identified in yeast to be required in the absence to the yeast DNA helicase SGS1, acts as an catalytic endonuclease complex with SLX1 and a scaffold for at least the other substrate-specific human nuclease Rad1-Rad10 (yeast XPF-ERCC1) and the similar MUS81-EME1 (Muñoz *et al.*, 2009; Rouse, 2009). With each of the three endonucleases showing specific preferences for types of strand breaks, the complex is thought to act as a 'molecular Swiss Army knife' to cover all aspects of needed repair (Cybulski & Howlett, 2011).

Additionally, the ICL homologous repair process involves a further main complex comprised several of the 15 Fanconi Anemia proteins (FA) and their associated proteins, so named for their involvement in the inherited, recessive genetic disease and the structure-specific nuclease FAN1 (MacKay *et al.*, 2010). During ICL repair, monoubiquitination of one of the FA proteins, FANCD2, recruits FAN1, which binds to the ubiquitinated Lys561 residue on FANCD2 via its UBZ-domain (MacKay *et al.*, 2010).

While the TONSL-MMS22L, Slx4 and FAN1 complexes appear crucial for DNA repair, understanding of the exact mechanics of the system and how each component is interlaced is nascent and remains to be further elucidated.

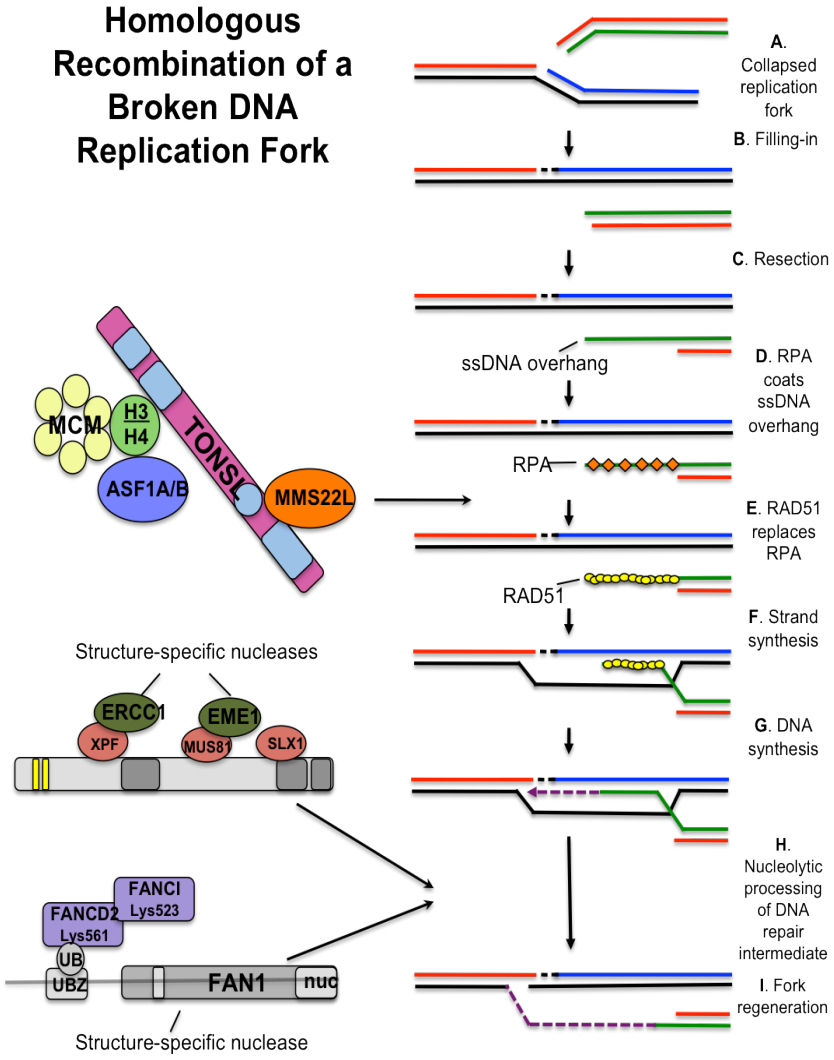


**Figure 5.1: Overiew of the current understanding of the Homologous DNA Repair Pathway.** The right side of the figure shows schematically the repair process as it occurs at the ICL damage site. The larger schematic on the left depicts the TONSL-MMS22L, SLX4 and FAN1 complexes at the present time with the arrows indicating where they are thought to be involved in the repair process. Figure adapted from the Rouse group website (http://www.ppu.mrc.ac.uk/research/?pid=7&sub1=research, accessed 28 August 2012).

### 5.1.2 SILAC Studies to Identify Protein Interactions

While co-immunoprecipitation experiments (described in more detail in Chapter 1.2.2.1: Immunoprecipitation and Co-Immunoprecipitation) are able to reliably identify complexes of proteins, they are hindered by non-specific binding that can result in noisy data and a large number of false positive interactions. As a response to this limitation, a recently developed extension of the method involving stable isotope labelling of amino acids in cell culture (or SILAC), a proteomics-based technique for detecting protein complexes based on changes observed in mass spectrometry profiles, has been proven a successful strategy for protein interaction (Trinkle-Mulcahy *et al.*, 2008; Boulon *et al.*, 2010; Westman & Lamond, 2011; Boisvert *et al.*, 2012). Mass spectrometry is frequently undertaken on the complexes returned from co-immunoprecipitation experiments due to the high sensitivity of protein detection that uncovers not only those entities bound directly to the target protein, but non-specific binding partners as well. Balancing this sensitivity with the specificity of interaction detection unfortunately means that the results obtained from mass spectrometry analysis are not always clear. However, SILAC offers a solution for differentiating between false positive hits and genuine interactions.

Figure 5.2 provides an outline of a double encoding SILAC experiment. In the first stage, a sample of cells containing a GFP-tagged target protein of interest is compared against a negative control sample containing the tag alone. The control and test cell populations are cultured in growth media supplemented with arginine and lysine amino acids with either their normal carbon (12C) atom (for the control population, red), or a combination of heavy carbon (13C) and heavy nitrogen (15N) atoms (yellow and blue).

After five growth cycles, cells in each of the populations have completely incorporated the carbon and nitrogen isotopes at no consequence to their viability and integrity. The cells are then lysed and mixed together into one sample that is purified through a variation of co-immunoprecipitation, in which complexes with the GFP-tagged protein covalently bind to GFP-interacting proteins coating Sepharose beads (green circles). The eluted complexes are then digested with trypsin, a serine protease that cleaves non-specifically at arginine and lysine amino acids, and the cleaved peptide complexes are analysed by mass spectrometry.

Amino acid replacement of the heavy carbon and nitrogen isotopes induces a shift in the molecular weight of the proteins that is readily detectable with mass spectrometry. Following quantification of the molecular weights of each protein complex retrieved from the co-immunoprecipitation, the ratio of heavy:light weights is calculated for each identified complex. The 'light' control population, which only includes the GFP tag itself, serves as a background control, such that a heavy:light ratio of 1:1 indicates that the protein in that sample is co-purified both with and without the target protein of interest. Therefore, the co-purified protein is likely to be either binding to the GFP tag on its own, to the Sepharose beads or non-specifically. Conversely, a higher heavy:light ratio indicates that the co-purified protein is binding only when the GFP-tagged target protein is also present, suggesting a specific interaction with the target protein.

While this double encoding form allows comparison of one cell condition to a control, SILAC experiments can also take a 'triple encoding' form in which two different cell or protein conditions (e.g. a wild-type and mutant protein or untreated and treated DNA) are simultaneously compared through labelling one of the test samples as a 'medium'

sample with 15C-arginine and the other as a 'heavy' sample with the 15C-arginine and 13N-lysine isotopes. In these three-way studies, the ratios of light:medium:heavy can be analysed against each other individually or as a whole. Similar to double encoding experiments, a 1:1:1 ratio between a co-purified protein suggests it is a non-specific interactor and higher ratios between either the medium:light or heavy:light indicating a genuine interaction in one or both of the assessed conditions.

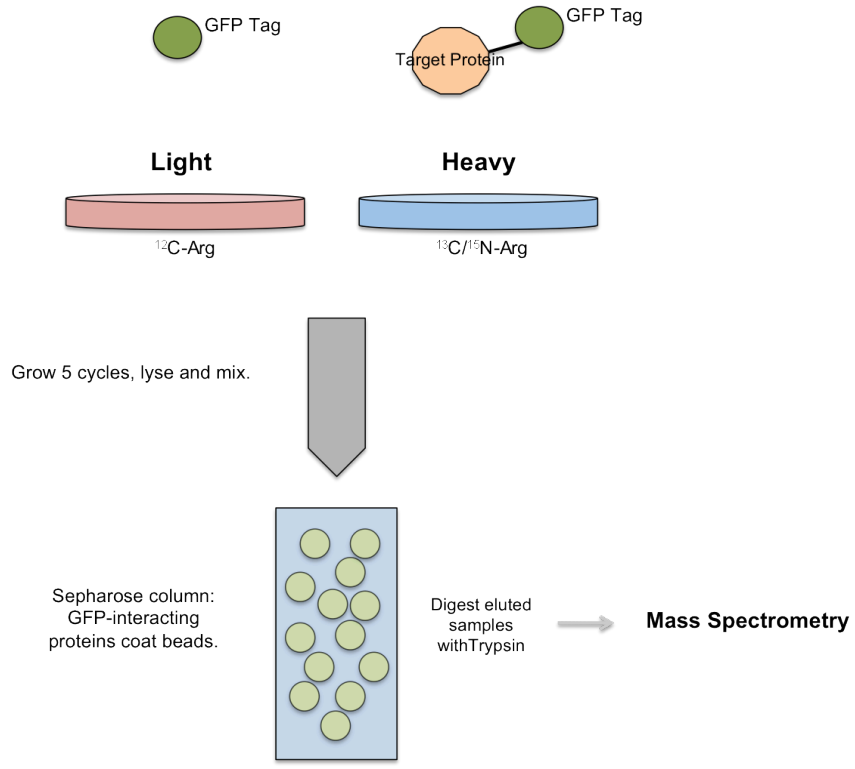


**Figure 5.2: Overview of the SILAC experimental protocol.** A target protein of interest (orange) is labelled with a GFP tag (dark green) and grown in three different culture media containing either light (red, a control) or heavy (blue) isotopes of arginine and lysine. After five growth cycles, the samples are mixed together and eluted through a column with Sepharose beads coated with GFP-interacting proteins. The eluted samples are digested with trypsin, which cleaves at lysine and arginine residues and resulting peptides analysed by mass spectrometry.

To analyse SILAC results, an arbitrary cut-off threshold is set for the heavy:light and, if applicable, medium:light ratios, that represents the decision point for a co-purified complex to be attributed to contamination or non-specifically binding of a protein, or a genuine interaction. However, setting an absolute limit has the potential to overlook genuine interactions that might have ratios that are not distinguishable from the background noise. Additionally, proteins with ratios above the threshold might only be non-specifically binding to the tag or beads. As a result, the Lamond Lab has added another dimension to this analysis by developing the Protein Frequency Library (PFL) (Boulon *et al.*, 2010). The PFL currently contains a record of the number of times the 30,366 proteins are detected across 185 SILAC-based experiments, and is incorporated by assessing the number of times that a protein appears in a range of experiments. If a protein is observed frequently, it is more likely than not to bind non-specifically or a be contaminant and not a genuine interactor in every complex it is seen in. Therefore, filtering SILAC results that fall above the cut-off threshold to exclude those belonging to proteins in the PFL that appear in more than a set percentage of experiments allows removal of further false positive interactions. However, while this filtering does allow the set of identified interactors to be narrowed down to those most likely to be true interactions, it still leaves large numbers of interactions for consideration.

Incorporating PIPs into these studies provides a further measure of filtering to suggest the most probable genuine interactions. Ideally, complexes with high-scoring M/L or H/L SILAC ratios would also score highly in PIPs and be predicted as interactions. More importantly, however, implementing PIPs into the SILAC analysis framework could provide additional insight into those complexes with low, but still above the threshold, ratios. While the majority of these complexes are likely to be insignificant

and not of interest, it is possible that some may be genuine interactors with low isotope ratios that should not be excluded.

#### 5.1.2.1 Cullin-4B (CUL4B)

Cullin-4B (CUL4B), a member of the cullin family of proteins, is a RING (Really Interesting New Gene)-type E3 ubiquitin ligase with an established role in ubiquitin-dependent protein degradation (reviewed in Sarikas *et al*., 2012). Like the other seven mammalian cullin proteins (CUL1, CUL2, CUL3, CUL4A, CUL4B, CUL5 and CUL7), CUL4B structurally consists of a stalk-like N-terminal domain marked by three cullin repeats and a C-terminal domain with the globular cullin homology domain. As a scaffold, the cullin proteins bind at least three additional and isoform-specific proteins: an adapter protein at the N-terminal domain, a signal recognition protein bond to the adapter protein and a RING-finger Zn-binding protein at the C-terminal domain, shown schematically in Figure 5.3, below.

At the C-terminal domain, the RING-domain protein Rbx1 (also called ROC1), recruits the E2 ubiquitin-conjugated enzyme to the complex. For its adapter protein, CUL4B is known to interact with the 127kDa adapter protein DDB1 (for damaged DNA binding protein), known to be involved in, at the very least, cell cycle regulation, cell death, transcriptional regulation and embryo development ((Lee & Zhou, 2007)). Structurally, DDB1 consists of three β-propellors, each with seven WD40-repeats (approximately 40 amino acids arranged into four anti-parallel β-sheets), arranged around a central axis and C-terminal domain (Li *et al.*, 2006), that allows for multiple contact points with a diverse range of substrates and a degree of flexibility to position those substrates in proximity to the C-terminal enzyme (Lee & Zhou, 2007; 2012).

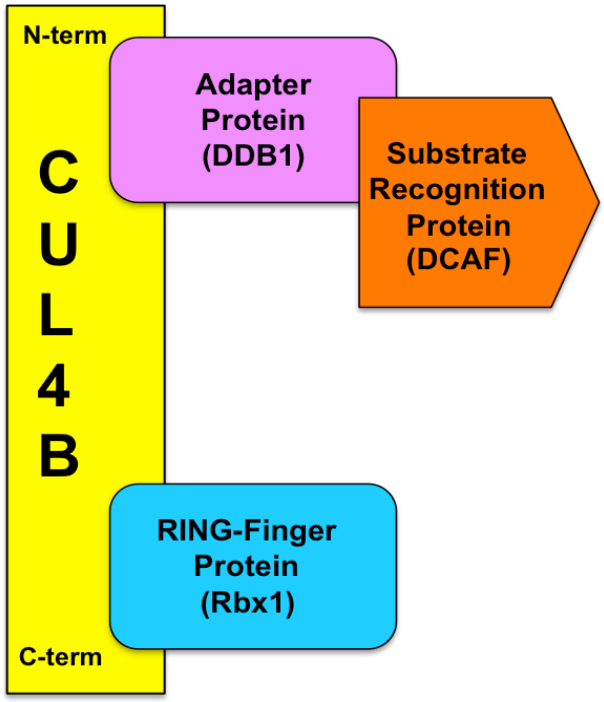


**Figure 5.3: Schematic of the CUL4B-DDB1-Rbx1 scaffold complex.** The N-terminal domain binds the adapter protein DDB1 that recruits and binds a diverse range of substrates through an associated DCAF (DDB1-CUL4-associated-factor). The C-terminal domain binds the RING-finger protein Rbx1, which then recruits the E2 ubiquitin-conjugating enzyme that catalyses degradation of the targeted substrate bound to the DCAF. Figure adapted from Sarikas et al., 2011.

Attempts at large scale identification of DDB1 substrates has identified at least 60 novel interactors, 56 of which contained a similar, one β-propellor, structure to DDB1, with diverse functional capabilities ranging across cell cycle and transcription regulation, cell signalling and chromatin modification among others (Higa *et al.*, 2006; Lee & Zhou, 2007). While achieving a full understanding of the extent of involvement of these DDB1-and-CUL4-associated-factors (DCAFs) is still in progress, evidence suggests that certain substrates may play more of a role than just a DDB1-CUL4 bridge and are likely helped by accessory proteins and cofactors and undergo varying degrees of modification that may regulate the recruitment and activation of the E2 enzyme.

While most research has centred around the CUL4A isoform, knock-out studies of CUL4A and CUL4B have shown them to be functionally redundant in most situations (Sarikas, Hartmann & Pan, 2011). However, mutations in CUL4B specifically have been shown to X-Linked Mental Retardation Syndrome (XLMR) (Tarpey *et al.*, 2007), possibly by altering the tight regulation of the DCAF WDR5, a member of the H3K4 methyltransferase complex, leading to increased expression of the protein and neuronal dendrite extension (Nakagawa & Xiong, 2011). After observation that this effect is unique to CUL4B and not CUL4A, despite an 80% sequence identity, it was discovered that only CUL4B contains a nuclear localisation sequence and is thus located, in its wild-type form, in the nucleus along with WDR5 (Nakagawa & Xiong, 2011). Additionally, recent studies have linked CUL4B to other substrates involved in specific processes affecting transcriptional regulation, DNA repair and cell cycle regulation (reviewed in Lee & Zhou, 2012). Finally, CUL4A and CUL4B have been shown to be involved in a downstream process of 'neddylation', in which the recruitment of the E2 ubiquitin-conjugating enzyme is increased and decreased through a cycle of binding between the ubiquitin-like NEDD8 and CAND1 proteins, respectively (Osaka *et al.*, 1998; Liu *et al.*, 2002; Pan *et al.*, 2004).

Though an increasingly strong image of the CUL4-DBB1 ubiquitinylation system is emerging, the rapidly growing number of DCAF substrates and functional implications of these interactions suggests that understanding is far from complete.

# 5.2 Methods

## 5.2.1 Prediction of the Interactions for Proteins in the Homologous DNA Repair System

### 5.2.1.1 Interaction Prediction and Results Presentation

Sixteen proteins known to be involved in homologous DNA repair, provided in Table 5.1, were selected for interaction prediction. Predictions from PIPs were returned for each of the three Bayesian predictors (EOCT, EOCM and EOCZ). In order to maximise the number of predictions returned as some of the proteins were not present in the PIPs database at the time, lower-than-normal (i.e. than 1000.0) cut-off thresholds were used for the prediction set.

The returned predictions were provided in a website accessible to the Rouse group, with the results for each protein of interest presented on a separate page that gave the final score for each of the three predictors and along with a 1-5 numerical ranking system to relay the contribution of each module to the final scores in a manner understandable to the group. Each table was sortable by the score columns to allow for ease of comparison between the high and low-scoring results. Additionally, each protein name included a link leading to its GeneCards (Safran *et al.*, 2010) page to allow easy access to further information from a range of sources about the prediction. Finally, any known interactions for the query protein were also returned and included the table marked as a known interactor.

| Query Protein | Full Name |
|---|---|
| EME2 | Probable crossover junction endonuclease EME2 |
| C9orf84 | Uncharacterised protein C9orf84 |
| RNF212 | RING finger protein 212 |
| ERCC4 | DNA repair endonuclease XPF |
| SLX4 | BTB/POZ domain-containing protein 12 |
| C1orf124 | Zinc finger RAD18 domain-containing protein C1orf124 |
| FAM60A | Tera protein homologue; FAM60A |
| TONSL | Tonsoku-like protein |
| ASF1A | Histone chaperone ASF1A |
| ASF1B | Histone chaperone ASF1B |
| C12orf48 | PARP1-binding protein |
| FANCD2 | Fanconi anaemia group D2 protein |
| FANCI | Fanconi anaemia group I protein |
| EME1 | Crossover junction endonuclease I |
| TERF2 | Telomeric repeat-binding factor 2 |
| TERF2IP | Telomeric repeat-binding factor 2-interaction protein 1 |

**Table 5.1: List of DNA repair proteins included in the prediction dataset.** Shortened and full names are given for each protein in the dataset.

As a method of facilitating the analysis of the predictions, the GO terms associated with each predicted interactor were also returned and included in each of the tables. An additional page provided grouped the predictions by each GO term such that results related to, for example, 'DNA Repair', could be easily identified from the rest of the predictions. In each table on the GO term page, a link was also provided to an antibody for the interactor, if available.

Results were then manually examined for predictions that seemed most credible based on their known or suspected molecular properties, nuclear localisation or biological function.

## 5.2.2 Incorporation of PIPs and PIP'NN with SILAC Studies in Nucleolar Proteins

### 5.2.2.1 Protein Dataset

Experimental data from co-immunoprecipitation and SILAC studies of the protein cullin-4B (CUL4B, UniProt Accession: Q13620), provided by the Lamond Group (University of Dundee, College of Life Sciences), were obtained with the procedure outlined briefly below (Boulon *et al.*, 2010).

Three populations of cells were grown in culture media enriched with different combinations of stable isotope-labelled arginine and lysine amino acids (see Table 5.2, below). After five to six splitting cycles, cells were extracted, GFP-tagged and incubated with an affinity matrix of Protein G-Sepharose beads coated with anti-GFP antibodies. After washing unbound molecules form the matrix, bound GFP-tagged-

protein-CUL4B complexes were eluted and mixed together with a 1:1:1 light to medium to heavy ratio of cell samples.

| Cell Population | Arginine Isotope Label | Lysine Isotope Label |
|---|---|---|
| Light (Control) | L-[$^{12}C_6$] arginine (R6) | --- |
| Medium (Bait, No Treatment) | --- | L-4,4,5,5-D4-lysine (K4) |
| Heavy (Bait, With Treatment) | L-[$^{13}C_6$,$^{15}N_4$] arginine (R10) | L-[$^{13}C_6$,$^{15}N_2$] lysine (K8) |

**Table 5.2: Details of arginine and lysine isotope labelling of the light, medium and heavy cell populations.** Isotope labels for the light (control), medium (bait with no treatment) and heavy (bait with treatment) cell populations are given.

Protein complexes were then digested into peptides with trypsin, extracted and separated with SDS-PAGE and then sent for liquid chromoatography-tandem mass spectrometry analysis. Mass spectrometry results were analysed with the MaxQuant software and the Mascot search engine to calculate ratios of medium:light (M/L) and heavy:light (H/L) for each identified SILAC pair.

#### 5.2.2.2 Prediction of Potential Interactors with Low and High SILAC Ratios

Complexes with M/L or H/L SILAC ratios above a selected threshold of 1.0 were included in the analysis dataset. The M/L and H/L datasets were kept separately, and each set was then split into two groups: a ‘low-ratio’ set with complexes with SILAC ratios between 1.0 and 5.0, and a ‘high-ratio’ set with complexes with SILAC ratios greater than 5.0.

In order to further analyse the set of complexes with M/L or H/L ratios at or around the cut-off threshold for evidence of genuine interaction, scores from both the PIPs EOCT predictor and PIP'NN were calculated for each CUL4B-protein pair. Results were then filtered at two levels. First, all pairs with EOCT scores less than the Bayesian PIPs prediction threshold of 1000.0 and PIP'NN scores less than the PIP'NN cut-off score for prediction of 0.5 were removed. In the second stage, both sets of predictions were considered against the Protein Frequency Library (PFL) (Boulon *et al.*, 2010) to remove any pairs in which the interacting protein was recorded in the resource to have a frequency of detection greater than 40%.

The remaining pairs predicted by PIPs and PIP'NN were then considered to identify possible interactions of interest. In order to identify pairs that were most probable, based on their biological classification, molecular function or subcellular localisation, to interact with CUL4B, the GO terms associated with each predicted interactor were downloaded from the PIPs database.

Complexes in the 'high-ratio' dataset with the highest-scoring M/L and H/L SILAC ratios were examined in two ways. First, the M/L and H/L sets were individually filtered with the same method as described above for the 'low-ratio' set. Second, the PIPs and PIP'NN scores and predictions were calculated for each 'high-ratio' pair without additional filtering. Predictions were then manually examined for interactions of interest based on their physical properties or known biological functions.

# 5.3 Results

## 5.3.1 Prediction of Protein Interactions in the DNA Repair System

Manually assessing the interactions predicted for each of the proteins of interest identified several interactors that both scored highly and were likely to be associated in some way with the homologous DNA repair system. As PIPs is built around prediction based on available evidence, a greater number of predicted interactions were returned for the proteins that have been more well-studied and characterised. Consequently, both the number and validity of predictions for the more established members of the system in the datasets (i.e. ERCC4, ASF1A and ASF1B) was greater than for those more recently discovered (i.e. SLX4, TONSL and C1orf24). Unfortunately, this is a limitation of an evidence-based method, particularly when applied to a biological system that is still largely not understood.

However, several of the predictions returned for five of the proteins in particular (FANCI, FANCD2, ERCC4, ASF1A and ASF1B) are plausible either based on their cell location, known involvement in the DNA repair process or known interaction with other members of the repair pathway. These predictions of interest are provided in Table 5.3 with their EOCT and EOCM prediction scores.

Several of the high-scoring predictions are known interactors, though they were not yet annotated in the version of PIPs implemented for the predictions. While these interactions do not offer insight into novel, potential interactions to test in the lab, they do bolster the capability of PIPs to predict interactions that it should with high values.

In particular, the predictor correctly identified the interaction between FANCD2 and FANCI known to be involved in the final stage of the homologous repair process with the recently discovered endonuclease FAN1. Additionally, the predictions for FANCI also identify RAD51API, the RAD51-associated protein, as a potential interactor. Given the clear involvement of RAD51 in the repair pathway, an interaction between the two proteins is not completely unlikely.

| Query Protein | Predicted Interactor | EOCT Score | EOCM Score |
|---|---|---|---|
| FANCI | FANCD2 | 3400 | 39.5 |
| FANCI | RAD51API | 3400 | 1423 |
| FANCI | BRCA2 | 32 | 1240 |
| FANCI | MCM6 | 1995 | 12 |
| FANCI | MCM10 | 1241 | 528 |
| ERCC4 | XPA | 311 | 11867 |
| ASF1A | HIST1H4A | 188529 | 3457 |
| ASF1A | ASF1B | 27866 | 190247 |
| SLX4 | TERF2 | 367 | 140 |
| SLX4 | TERF2IP | 1709 | 65 |
| SLX4 | RAD54L2 | 36 | 0.4 |
| FANCD2 | MCM3 | 2866 | 463 |

**Table 5.3: Predicted Interactions of Interest in the DNA Repair System.** A selection of interactions predicted by both the EOCT and EOCM PIPs predictors that are either known (highlighted in grey) or plausible interactions based on their known biological functions or established role in the repair pathway is provided. Scores are given before adjustment for the prior odds ratio (i.e. before dividing by 1000) for ease of comparison. For a true assessment of the likelihood that the two proteins will interact, the scores above should, therefore, be divided by 1000.

Several of the MCM (mitochondrial maintenance complex component) proteins were also predicted to interact with FANCI and FANCD2. While the MCM proteins may not interact directly with FANCI, they are known to interact with the ASF1A and ASF1B histone-binding proteins that join the TONSL-MMS2L complex during the repair process. Analysis of the contributions of the individual modules to the final prediction score revealed that the majority of the highest-scoring predictions for the FANCI and FANCD2 proteins were due to either or both high scores in for gene co-expression or the transitive module, both of which reflect potential involvement in a similar pathway or network.

For the less-characterised proteins, in particular SLX4, the highest scoring predictions were several tens of times lower than the highest scoring for the more well-known examples in the dataset. While this drop in scores required a decrease in the cut-off threshold to identify some interactions, several of the predictions did include known members of the DNA repair system. For example, with this lowered threshold, SLX4, TERF2 and TERF2IP were all predicted as interactors of each other. While taking a hard approach and limiting 'predictions' to only those strictly above the 1000.0 threshold would eliminate these interactions, as they were included in the top hits predicted by PIPs for those proteins, they should not be discounted.

Analysis of the scores for known interactions for each of the query proteins, if applicable, that were included in the PIPs database did not score as expected. Assessing the breakdown of the module contributions showed that the low scores were due to the low likelihood ratios across all modules, suggesting either that the data in PIPs for the

two proteins does not strongly support interaction or there is not enough data available for one or both of the proteins to substantiate a positive prediction.

## 5.3.2 Identification of Potential Interactors with Low SILAC M/L and H/L Ratios

### 5.3.2.1 Dataset Selection

To determine the cut-off thresholds for splitting the identified complexes into low- and high-ratios, the M/L and H/L SILAC ratios were plotted as a histogram showing the distribution of ratios for all of the complexes (Figure 5.4). While the range of scores for both the M/L and H/L ratios vary between the complexes, as indicated by the non-uniform pattern of bars Figure 5.4 and the distribution of ratios in Figure 5.5, the majority for each set fall below 1.0, with many complexes having ratios slightly higher and only a few with clear, outstanding ratios. To give an indication of the variation between the ratios at or around the threshold, the insets in Figure 5.4 offer a closer view of the profile of lowest ratios observed. While each of these ratios could be considered 'low' in comparison with the complexes with the highest ratios, there are still complexes with ratios that could go either way to being attributed to background noise or caused by weakly expressed or low-signal complexes that are genuine interactions.

After assessing the score distributions for the M/L and H/L sets, a lower limit cut-off threshold of 1.0 and upper limit of 5.0 were selected for the 'low-scoring' dataset of complexes with ratios at or around the threshold.

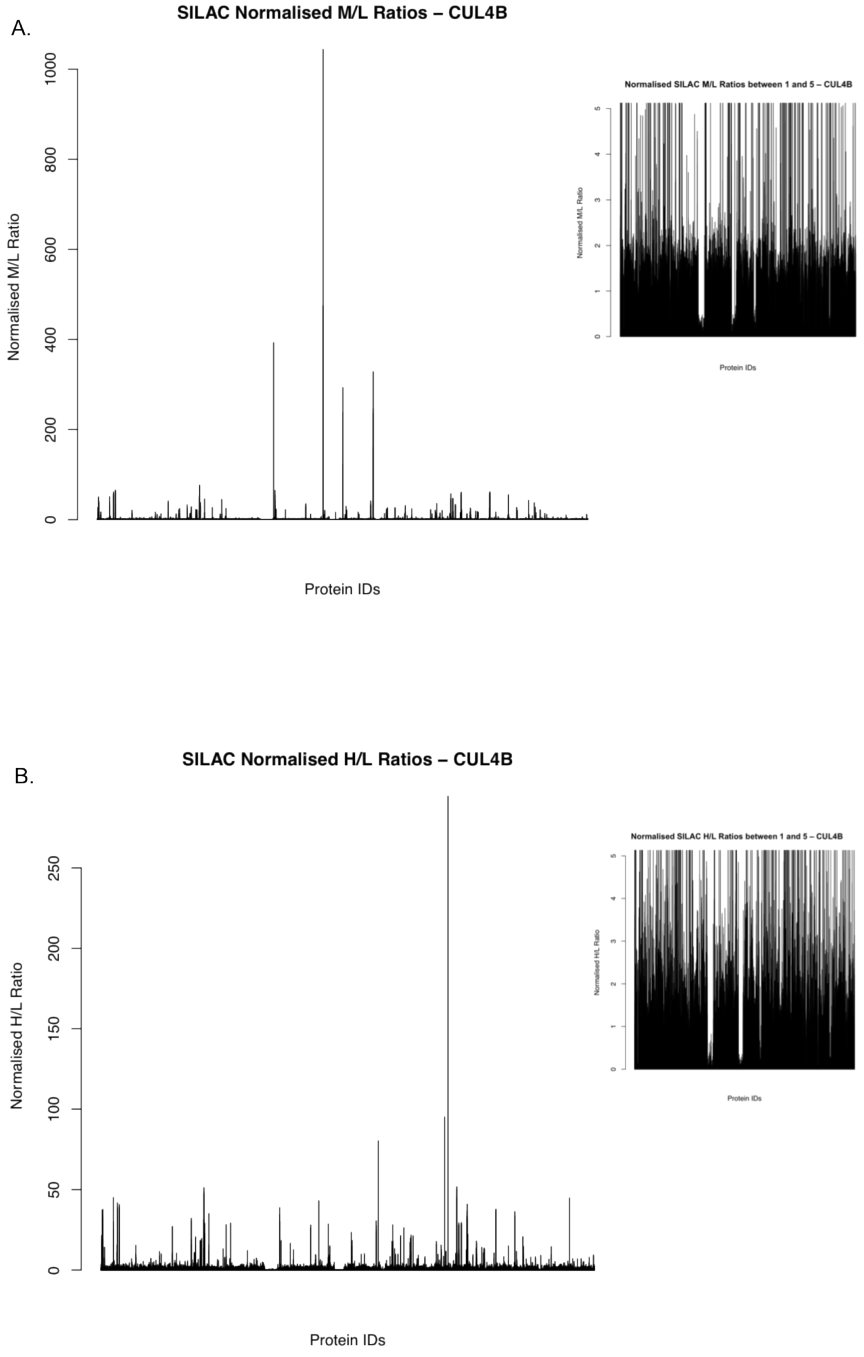


**Figure 5.4: Barplots of the M/L and H/L SILAC ratios for each identified protein complex.** (A) Normalised SILAC M/L ratios for the CUL4B experiment for each protein complex identified. (B) Normalised SILAC H/L ratios for the CUL4B experiment for each protein complex identified. The insets for (A) and (B) show in more detail the spread of interactions around the lowest ratios.

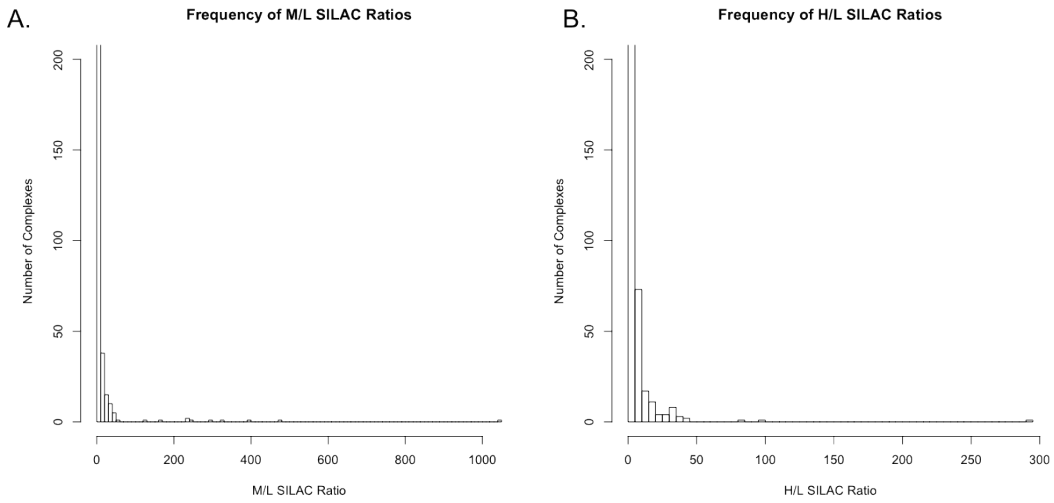


**Figure 5.5: Distribution of M/L and H/L SILAC ratios for CUL4B.** (A) Histogram of SILAC M/L ratios for CUL4B experiment. (B) Histogram of SILAC H/L ratios for CUL4B experiment. In both plots, the y-axis has been prematurely truncated at 200 to better show the distribution of scores with higher SILAC ratios. Actual counts for low ratios are above 10,000.

Predictions were made for each complex in the M/L and H/L datasets were then calculated by both the PIPs and PIP'NN predictors. Table 5.4 shows the number of complexes with PIPs and/or PIP'NN scores above the respective 0.5 and 1000.0 cut-off thresholds.

| | M/L | H/L |
|---|---|---|
| **Total Pairs** (Ratio > 1.0 and < 5.0) | 543 | 480 |
| **PIPs** (EOCT LR > 1000.0) | 2 | 2 |
| **PIP'NN** (Output Score > 0.5) | 100 | 86 |
| **Both** | 2 | 2 |

**Table 5.4: Number of Complexes with Interactions Predicted by PIPs and PIP'NN.** PIPs scores above the 1.0 cut-off threshold and PIP'NN scores above the 0.5 prediction threshold for the M/L and H/L low-scoring datasets.

#### 5.3.2.2 Potential CUL4B Interactors

In order to identify which, if any, of these complexes were most likely to be genuine interactions, the filtered results were then manually analysed for shared properties or evidence of any biological or functional similarity that might support an interaction. Table 5.5, below, details 17 of these complexes, chosen for either their high PIPs or PIP'NN scores, their molecular properties or interactions with proteins known to interact with the cullin proteins.

| Uni-Prot ID | Protein Name | M/L SILAC Ratio | H/L SILAC Ratio | PIPs EOCT Score | PIP'NN Score | Brief Description |
|---|---|---|---|---|---|---|
| P06493 | CDC2 | 1.4031 | 1.484 | 1.46331 | 0.859948 | Cyclin-dependent Ser/Thr kinase; phosphorylation to regulate G1/S and G2/M phase transitions |
| P62277 | RPS13 | 1.9481 | 2.113 | 0.655852 | 0.51176 | Ribosomal protein (small subunit) |
| P25787 | PSMA2 | 1.1176 | 0.944 | 160.245 | 0.55705 | Proteasome subunit alpha type 2; member of proteinase complex; ATP-dependent cleavage at R, F, Y, L and E residues |
| Q7L2H7 | EIF3M | 1.699 | 1.4662 | 2522.96 | 0.60801 | Eukaryotic initiation factor 3M; translation initiation and post-translational ribosomal disassembly |
| P52292 | KPNA2 | 1.6218 | 1.317 | 2969.31 | 0.540371 | Karyopheria alpha-2; nuclear import regulation by interacting with NLS of DNA helicase Q1 and SV40 T antigen; W(D)J recombination |
| Q14103 | HNRNPD | 1.6456 | 1.303 | 38.2141 | 0.996988 | Heterogenous nuclear ribonucleoprotein D; associates with pre-mRNAs to mediate mRNA stability; involved in cytoplasmic deadenylation and translational decay of FOS mRNA |
| P42766 | RPL35 | 1.902 | 2.009 | 1.24334 | 0.508346 | Ribosomal protein (large subunit) |

| Uni-Prot ID | Protein Name | M/L SILAC Ratio | H/L SILAC Ratio | PIPs EOCT Score | PIP'NN Score | Brief Description |
|---|---|---|---|---|---|---|
| P78527 | PRKDC | 2.1564 | 0.672 | 58.6248 | 0.720972 | Ser/Thr protein kinase; molecular sensor of DNA damage; involved in non-homologous 3'-end joining in DSB or V(W)J repair by protecting broken ends of DNA; Scaffold proteins; Phosphorylates histones H2AX, H2AFX and H1 |
| P62333 | PSMC6 | 1.8141 | 1.205 | 473.117 | 0.570016 | 26S protease; ATP-dependent degradation of ubiquinated proteins; part of 20S proteasome and PA700 complex |
| Q9H0S4 | DDX47 | 1.2664 | 0.900 | 1.24334 | 0.511438 | DEAD-box polypeptide 47; Apoptosis; Possible role in rRNA processing and mRNA splicing |
| P12004 | PCNA | 2.1981 | 1.891 | 3.19524 | 0.582503 | Proliferating cell nuclear antigen; DNA polymerase delta auxiliary protein that increases ability of polymerase for elongation; interacts with APEX2 with misincorporation of uracil |
| Q8IWA0 | WDR75 | 1.2743 | 0.877 | 19.383 | 0.547092 | WD-repeat 75; Contains WD-repeat |
| Q6PL18 | ATAD2 | 1.058 | 1.2149 | 58.6248 | 0.569306 | ATPase family - AAA domain containing 2; Transcriptional coactivator of ESR1 nuclear receptor to induce estradiol target genes; possibly involved in histone hyperacetylation |
| Q14566 | MCM6 | 1.896 | 1.2713 | 854.944 | 0.58651 | Minichromosome maintenance complex component 6; component of MCM complex; DNA helicase activity; might be involved in DNA unwinding and replication |
| O75717 | WDHD1 | 1.707 | 2.1158 | 49.8122 | 0.53974 | WD-repeat and HMG-box DNA binding protein; Replication initiation factor to associated MCM2-7 helicase and DNA polymerase for replication initiation |
| Q9NYL9 | TMOD3 | 1.165 | 4.0784 | 1.46331 | 0.538228 | Tropomodulin-3; Blocks elongation and depolymerisation of actin filaments |

| Uni-Prot ID | Protein Name | M/L SILAC Ratio | H/L SILAC Ratio | PIPs EOCT Score | PIP'NN Score | Brief Description |
|---|---|---|---|---|---|---|
| O43684 | BUB3 | 1.779 | 1.808 | 58.625 | 0.616 | Spindle-assembly checkpoint signalling and kinteochore-microtubule attachments; Role in inhibiting anaphase-promoting complex and ubiquitin ligase activity of APC/C with phosphorylation of CDC20; Phosphorylates MAD1L1 |
| Q9ULV4 | CORO1C | 1.570 | 1.249 | 22.812 | 0.547 | Coronin, actin-binding protein; Potential role in cytokinesis, motility and signal transduction |

**Table 5.5: Predicted interactions of possible interest.** Details of the UniProt identifier, common gene name, M/L and H/L SILAC ratios, PIPs score, PIP'NN score and brief notes on what is known about 17 selected interactors from the CUL4B SILAC experiments. Descriptions of functions taken from the UniProtKB entry for each protein.

Functionally, the predicted interactors highlighted are diverse; however, this follows the equally varied biological mechanisms and involvements of the already identified substrates for the CUL4B-DDB1 complex. Of these interactions, several are of particular interest. Four of these proteins, BUB3, WDHD1, CORO1C and WDR75, contain one or more WD40 or WD-repeat domains characteristic of the DCAF substrate recognition proteins that form part of the CUL4B-DDB1 complex, suggesting evidence for interaction based on their structural properties. Additionally, while the PIPs EOCT likelihood ratio is far below the 1000.0 cut-off threshold, that it is not overly low suggests that there is evidence in at least one of the modules supporting interaction. Looking at the evidence breakdown, each of these predictions appears based on recognition of the WD40-domain and a moderate to moderately-high gene expression correlation.

Between the two interactions of the entire filtered set were predicted by the PIPs predictor, the CUL4B-KPNA2 interaction is of particular interest. While the CUL4A and CUL4B isoforms share 80% sequence identity, they differ in CUL4B's inclusion of a nuclear localisation signal (NLS) that specifically targets it to the nucleus (Nakagawa & Xiong, 2011). With its dual functionality as a regulator of nuclear import and of W(D)J repair that places it in subcellular proximity to CUL4B and as a piece of the DNA repair process, it is possible that CUL4B might play some role in in KPNA2 regulation.

#### 5.3.2.3 Prediction Scores for Complexes with Highest M/L and H/L SILAC Ratios

In order to assess how PIPs and PIP'NN classified the complexes with the highest M/L and H/L SILAC ratios that most likely represent genuine CUL4B interactions, the PIPs EOCT likelihood ratio and PIP'NN output score for each interactions with ratios above 5.0 (the cut-off for the low-ratio dataset) were obtained. Table 5.6, below, details 12 of these identified interactors with the highest M/L and/or H/L ratio.

Of these interactions, only two - TCPQ, a chaperonin thought to assist with protein folding, and MYO1C, the 1C isoform involved in transcriptional regulation of the myosin ATP-powered motor protein - were predicted by PIP'NN. Examining the feature contributions for each of these interactions revealed that the lack of positive predictions were due to low scores across the modules in PIPs, which could be attributed to either a lack of evidence suggesting interaction or a lack of evidence available in general. Interestingly, however, the functions of these complexes, though again, diverse, could offer an additional level of support for some of the interactions

predicted for the complexes in the low-ratio dataset. For example, that two of the high-ratio interactions, one of which was also predicted by PIP'NN, are involved in some form of microtubule movement or cytoskeleton maintenance (CKAP4 and MYO1C), goes in line with the low-ratio predictions of TMOD3, BUB3 and CORO1C as interactors.

| Uni-Prot ID | Protein Name | M/L SILAC Ratio | H/L SILAC Ratio | PIPs EOCT Score | PIP'NN Score | Brief Description |
|---|---|---|---|---|---|---|
| Q6NUT2 | ABCB8 | 1043.7 | 28.581 | 0.983475 | 0.228391 | ATP-binding cassette, sub-family B |
| Q96PK6 | RBM14 | 328.08 | 1.300 | 1.46331 | 0.488704 | RNA binding motif protein 14; Nuclear receptor coactivator; Transcriptional repressor |
| Q15046 | SYK | 292.76 | 0.271 | 0.320362 | 0.378923 | Spleen tyrosine kinase; Two SH2 domains bind ITAMs for activation and autophosphorylation |
| P62877 | RBX1 | 50.392 | 37.257 | 3.4647 | 0.437398 | Ring-box 1; E3 ubiquitin protein ligase; Component of cullin-RING-based E3 ubiquitination complexes |
| P11021 | HSPA5 | 42.79 | 2.638 | 0.655852 | 0.479985 | Heat shock 70kDa protein 5/Glucose-regulated protein GRP78); In ER lumen; Helps with folding and assembly of proteins and protein transport |
| Q15843 | NEDD8 | 35.934 | 0.935 | 4.07766 | 0.450507 | Neural precursor cell expressed; ubiquitin-like; Involved in cell cycle control; Activates cullins when binds |
| P50990 | TCPQ | 26.884 | 0.1795 | 1.24334 | 0.523947 | CCT8L2; Chaperonin containing TCP1; Assists with folding |
| O75436 | VPS26A | 24.622 | 0.889 | 5.01745 | 0.411145 | Vacuolar protein sorting 26 homologue; Retromer complex (retrieves lysomal receptors from endosomes) |

| Uni-Prot ID | Protein Name | M/L SILAC Ratio | H/L SILAC Ratio | PIPs EOCT Score | PIP'NN Score | Brief Description |
|---|---|---|---|---|---|---|
| Q9BTV4 | TMEM43 | 22.009 | 18.564 | 0.320362 | 0.378043 | Transmembrane protein 43; Nuclear envelope structure maintenance |
| Q8NBN7 | RDH13 | 20.329 | 294.47 | 0.796492 | 0.233256 | Retinol dehydrogenase 13; No retinol dehydrogenase activity exhibited |
| Q07065 | CKAP4 | 1.7569 | 44.747 | 0.771883 | 0.447604 | Cytoskeleton-associated protein 4; Anchors ER to microtubules |
| O00159 | MYO1C | 6.504 | 7.853 | 0.244 | 0.604 | Myosin-IC; Transcription regulation with WICH chromatin-remodelling complex |

**Table 5.6: PIPs and PIP'NN scores for interactors with highest M/L and H/L SILAC ratios.** Details are given for twelve interactors with the highest M/L and/or H/L SILAC ratios and gives their UniProt ID, common gene name, M/L and H/L SILAC ratios, PIPs score, PIP'NN score and brief notes about what is currently known about the protein. Of these proteins, only two (highlighted in light grey) were predicted to interact by either PIPs or PIP'NN.

## 5.4 Discussion

Both of the investigations described offer practical implementations of the PIPs and PIP'NN predictors into lab experimental pipelines. While a machine-learning prediction technique will never, on its own, be able to replace in vitro and in vivo experimental identification of interactions, it can provide an initial suggestion for interactions that are worth considering, for one reason or another, as part of an experimental dataset.

While useful in this regard, prediction of interactions within a specific system can be limited by the availability of known data about proteins within that system. With the goal of most research to uncover novel information, often times experimental interests stand at the forefront of a field where little is publicly available about the exact topic being studied. This lack of evidence availability was particularly apparent within the predictions of DNA repair proteins for the Rouse group. While knowledge of the DNA repair system has been rapidly expanding over recent years, several of the proteins within their initial dataset for prediction were not yet present within the PIPs database or did not have any additional information available for any of the evidence features required for prediction. As a result, for some of the proteins that were in PIPs, the number of interactions with PIPs likelihood ratios above the 1000.0 cut-off was low or non-existent. While including predictions with lower final ratios into the result sets did increase the number of the interactions for consideration, in strict terms of what the final PIPs score indicates, these would not have otherwise been considered predictions. Regardless, several of the results returned did appear to be plausible potential interactions.

The incorporation of PIPs and PIP'NN into the analysis of SILAC results with the Lamond lab offers an additional way that protein interaction predictions could be implemented as part of a lab protocol. Adding this extra layer of filtering for complexes identified for CUL4B with low SILAC ratios did provide several examples of CUL4B interactions that might offer insight into novel substrates and implications for the complex. However, it is also possible that they are predicted based mostly on the gene expression feature that, on its own, forms only one piece to the puzzle of interaction. While correlated patterns of expression do support involvement in similar, simultaneous cell processes, they do not shed light specifically on whether the two proteins are involved in the same cell process. Therefore, while these predictions do offer a suggestion of a set of protein complexes that would otherwise be ignored based on their low SILAC ratios, true assessment of their reliability and validity will rely on further, specific experimental testing.

## 5.5 Conclusions

1) PIPs and/or PIP'NN can be incorporated practically into lab-based experimental protocols as either an initial means of narrowing down a dataset for experimental validation or filtering a set of results for those most likely to interact.

2) While several potential interactors were identified for the proteins in the DNA repair system, experimental validation of the most likely did not reveal any genuine interactions.

3) Practical use of machine learning techniques is largely dependent on how much evidence is available for the features considered by the predictor, limiting prediction on systems with little known and publicly available data.

4) Prediction of interactors of CUL4B with low SILAC ratios did reveal several proteins with either structural or functional similarity to known substrates of the complex as a set of potential novel substrates.

# Chapter 6

## Updates to the PIPs Web Server

## Preface

This chapter describes work done on the PIPs web server. Details of the current status of the server database, updates made to the framework and redesign of the site are provided along with suggestions for future maintenance and development of public access of the tool.

# 6.1 Introduction

While both PIPs and PIP'NN predictors are developed locally, the true merit of the tools lies in their ability to be accessed by the public for practical use. Although the PIPs v. 1.0 has been publicly available at http://www.compbio.dundee.ac.uk/www-pips since 2009 (McDowall, Scott & Barton, 2009), the data included in the web server database, including the set of predictions, has not been updated since its inception. Additionally, the current server was written in Java Servlet Pages (JSP) and styled with the HTML table format, technologies which are not only difficult to maintain without extensive knowledge of the web framework, but are also moving rapidly out-of-date.

Updating and maintaining the PIPs web server is therefore critical to allowing the work completed over the past three years to be made accessible in the public domain for continued use outside of the collaborations within the University of Dundee.

# 6.2 Updates to the Web Server

## 6.2.1 Development Framework

Updating the current web server framework could have required either a straightforward update to the data in the database and an alteration of any queries required to fetch the desired output; however, when coupled with the need to redesign the site styling as well, a decision was made to capitalise on the work already put into the site but rewrite it in its entirety. While JSP works well in conjugate with the Java-based, standalone PIPs framework, its lack of flexibility does not allow CSS styling, Javascript and more extensive layouts to be added easily. Several approaches were considered for the new

framework, for example the Python web-database framework Django or the Ruby framework Ruby on Rails, both of which are rapidly increasing in popularity of use in the web development arena. However, while both of these frameworks are ideal for dynamically and frequently updated underlying databases with user input/output, the relatively static nature that the PIPs website requires negates the need for such a heavy backend structure. As the PIPs website is built around a series of straightforward queries to the web server MySQL database, the PHP scripting language was selected instead. While PHP does have many downsides, namely its lack of robust security and its somewhat older age, for interfacing with MySQL and converting output into styled HTML, it was adequate enough for what was needed. As a result, the website is written in a combination of HTML5 interlaced with the PHP 5.4 stable version (5.4.3).

An updated version of the Nucleolar database forms the backend data server for the new web server and includes any new data incorporated into PIPs v. 3.0 and the set of new interaction predictions. For speed of data recovery, as was done in the previous database, interaction predictions above the PIPs cut-off thresholds of 1.0, 10.0 and 100.0 are stored separately in a table from the rest of the negative predictions and indexed appropriately. This separation dramatically reduced the size of the main interaction prediction table to 601,437 interactions with final PIPs EOCT, EOCM and EOCZ scores above 1.0 and increased the required to query the site and populate the main results page.

The website workflow follows the same general pattern as the original site. The site homepage, shown in Figure 7.1, contains a brief summary of the predictor, links to the citations for PIPs v. 1.0 and its web server, and two simple search forms. The first

search option allows a simple search for the predicted interactions for a protein by its UniProt, IPI or Ensembl identifier at a choice of thresholds (1.0, 10.0 or 100.0) that are provided in a drop-down selection box. The second search option allows a text search by word or phrase. A link to the EBI's PICR protein identification mapping service and the UniProtKB ID mapping service are also provided for easy reference for one of the accepted identifiers. Additionally, a link for the advanced search page is provided both by the search fields and in the navigation bar at the top of the page. Currently, advanced searching allows for the option of submitting a batch list of identifiers separated by new lines, searching by the gene name of the protein or searching by sequence. The sequence search function is a local BLAST search against the PIPs database of proteins that returns that top matches for the query protein. Each of the advanced search options (by text search, gene name or sequence) then redirects to a new page listing the matched proteins in the database that link to the set of interaction predictions and then number of predictions above minimum scores of 1.0, 10.0 and 100.0, that each link to the main results page.

Results are returned on a main results page, shown in Figure 7.2, that includes easily accessible links to information about the query protein and a sortable table with the resulting predictions. Additionally, the table includes four columns with a coloured circle corresponding to each source of evidence for that pair analysed by the predictors and indicating its contribution to the overall prediction and a link to a specific 'evidence' page for each prediction. Finally, the table contains a column indicating if the predicted interaction is already known and annotated in either the I2D, STRING, DIP or HPRD databases.

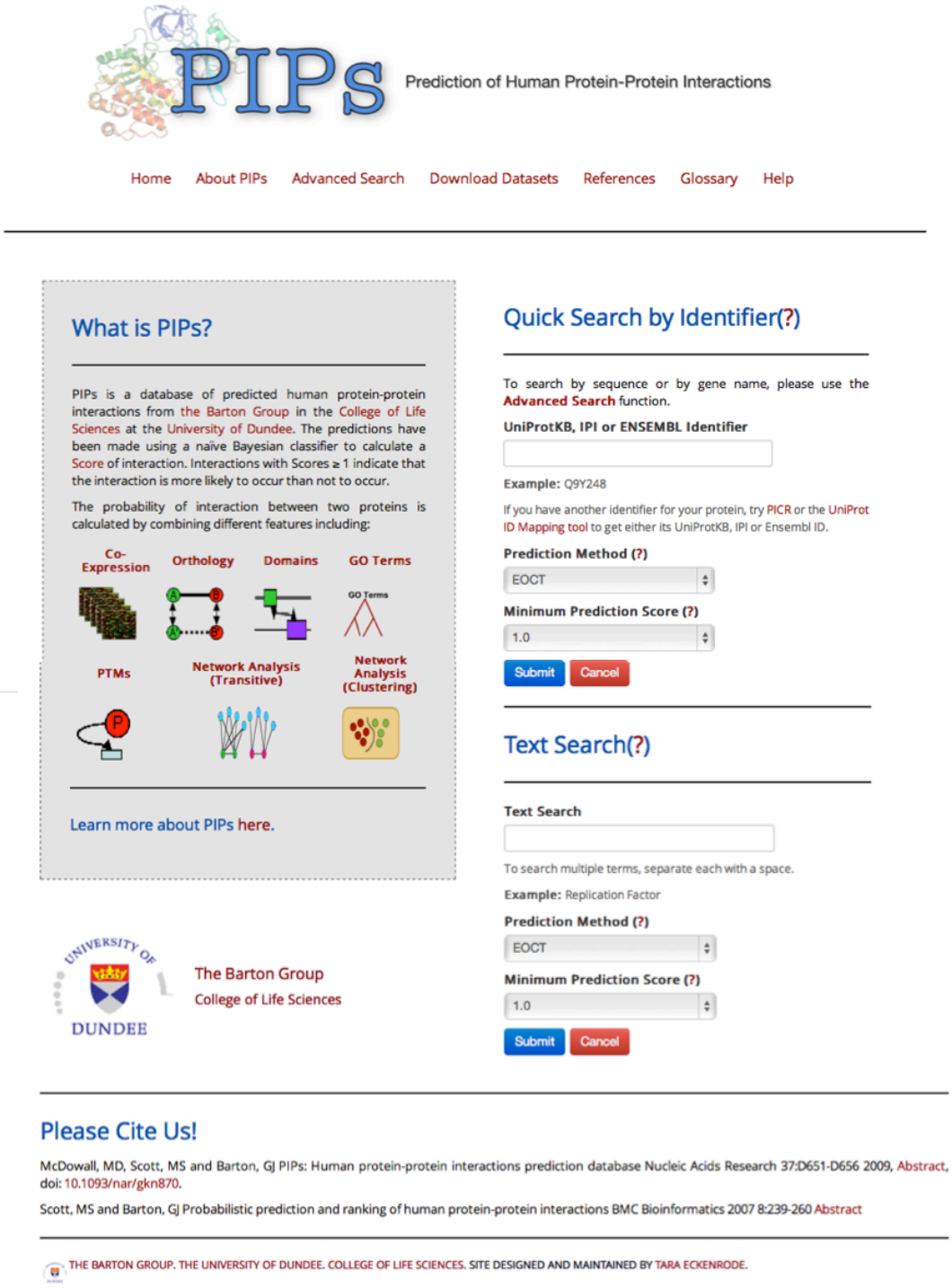


**Figure 7.1: PIPs homepage.** The screenshot above of the PIPs homepage shows the initial prediction form allowing the user to enter the UniProtKB, IPI or ENSEMBL identifier for their protein of interest and select a threshold from a dropdown list of options.

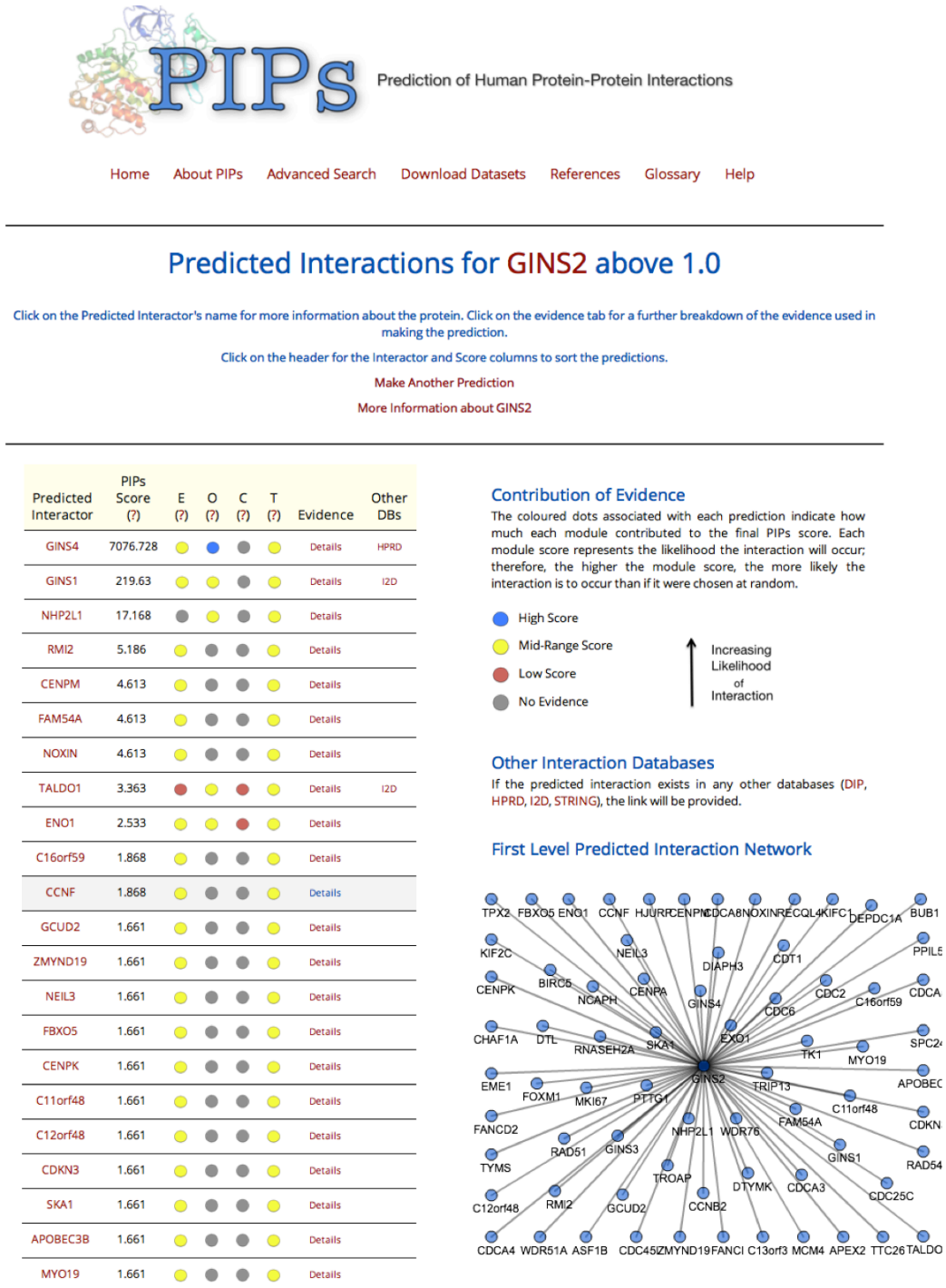


**Figure 7.2: Main query results page. Screenshot shows the main results page for a query for the protein GINS2.** The table on the left side of the main section lists the names of the predicted interactors, the PIPs score and then provides a colour-coded circle for each of the modules that represents how much that module or feature has contributed to the final prediction score. Additionally, a link to 'Details' and, if applicable, links to other databases where that interaction is recorded are provided. At the top of the page, there are also links to 'Make Another Prediction' or 'More Information About GINS2.

Clicking on the 'Details' link for each predicted interaction leads to a page with six tabbed sub-pages ('Expression', 'Orthology', 'Combined', 'Transitive', 'More Information'). The 'Expression', 'Orthology', 'Combined' and 'Transitive' pages each include a breakdown of the evidence in each module incorporated into the final prediction, along with its individual module score.

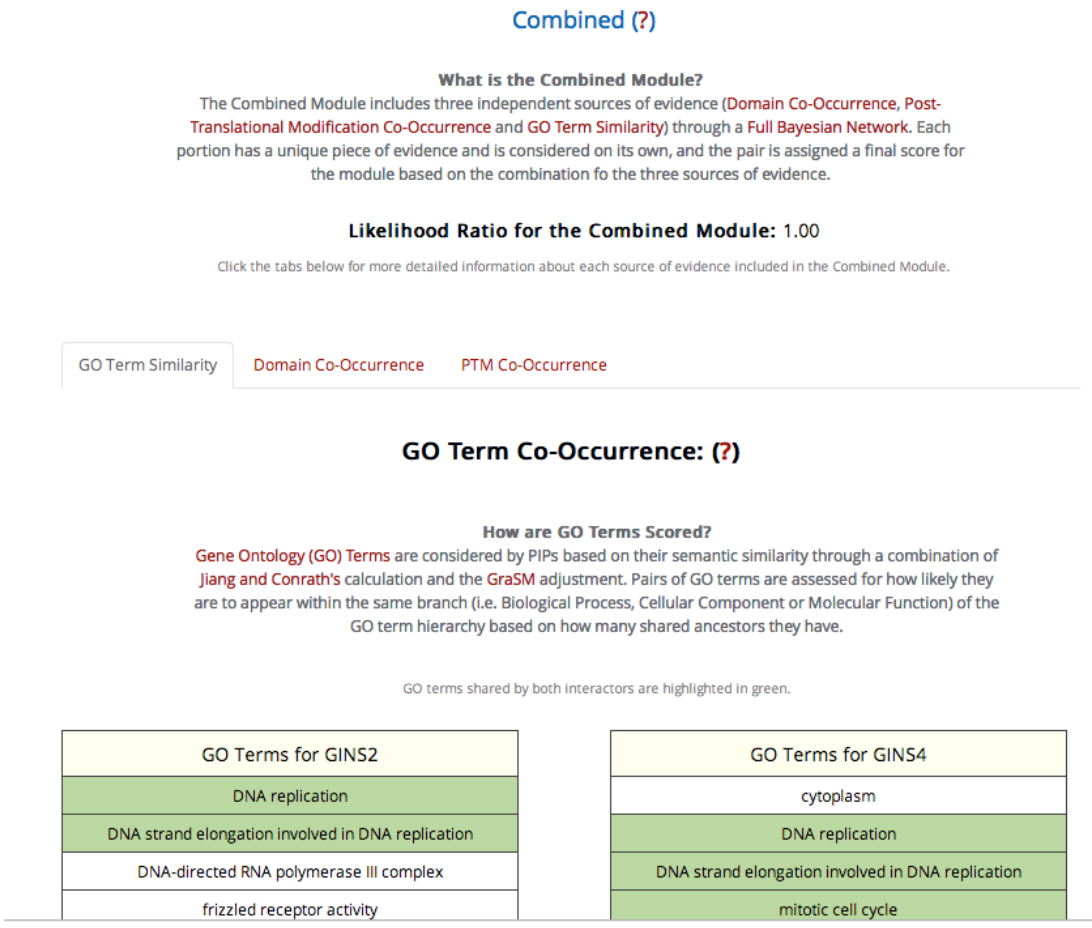


**Figure 7.3: Combined module evidence page.** The above screenshot shows the evidence tab for GO term portion of the Combined module. For each source of evidence, the score for that module is provided, along with, if applicable, a detailed breakdown of the specific evidence incorporated into calculating that score. For the Orthology and Combined modules, the features for each protein in the pair are displayed in side-by-side tables with similar/identical features highlighted in light green for ease of identification.

The ‘Orthology’ and ‘Combined’ pages also include tables showing the feature considered for each protein in the pairs with overlapping features highlighted (for example in the ‘Combined-GO Terms’ page, shown in Figure 7.3, all GO terms of each protein are provided, and any shared GO terms are highlighted in green). The ‘More Info’ tab, shown in Figure 7.4, contains information about the predicted interactor, including its name, other known identifiers, its sequence, numbers of interactions predicted above the thresholds of 1, 10, 100 and 1000 in PIPs and any known interactions contained in the database.

Additionally, there is a similar style tabbed page containing details about the query protein that includes the same information as on the ‘More Info’ page for the predicted interactor. Finally, the site contains several static pages including about, glossary and help pages, a page containing links to downloadable files with predictions at certain thresholds, and a page with citations for resources that may be of interest to protein interaction prediction.

The website has been styled with Twitter Bootstrap (http://twitter.github.com/bootstrap/), a pre-packaged set of CSS and grid-style layouts and Javascript functions that allows easy site design and theme modification. Building off of this template, the site is styled with to have a modern, clean and intuitive feel. For example, the use of tabbed sub-pages to house the more detailed information contributes to giving the site an uncluttered look without resources being obscurely hidden.

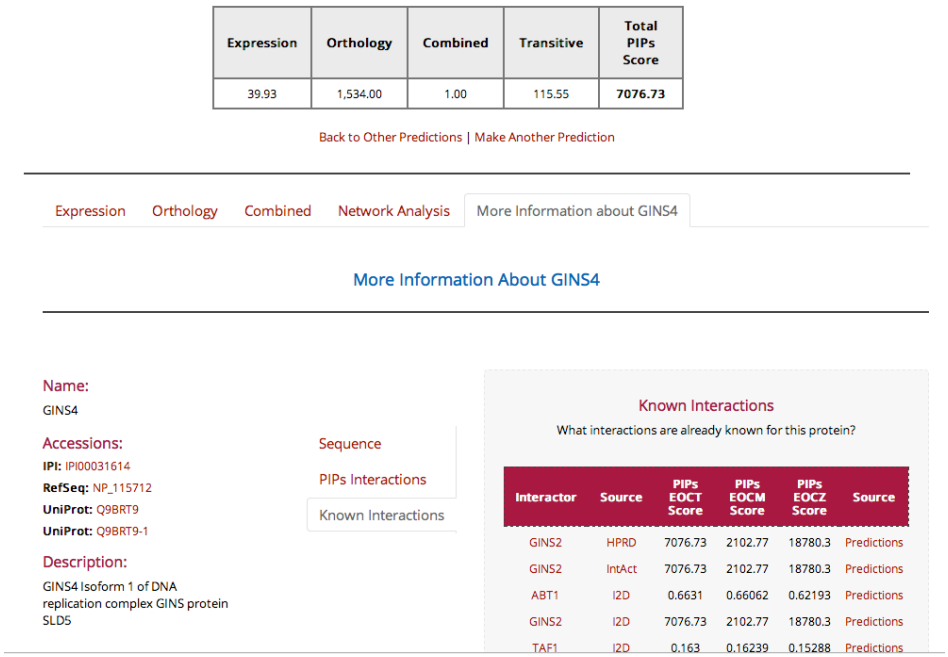


**Figure 7.4: Example of 'More information' page for the predicted interactor.** Screenshot shows the layout of the 'More information about' page for the predicted interactor when on the 'Known Interactions' tab. The left column remains static and displays the name of the protein, any other accession IDs it has and a brief description of its name. The right side tabs are clickable and open up two other pages with the amino acid sequence of the protein and details of how many interactions are predicted by PIPs at different thresholds.

Overall, the functionality of the original PIPs website has been maintained. Where possible, database queries that had been optimised for efficient data retrieval were kept as they were to maintain site performance. HTML and PHP code is all documented both inside and outside of the scripts to allow the site to be updated and fixed if and when it is necessary.

## 6.3 Future Directions

In the future, the web server could be adjusted further to incorporate more network analysis. While currently, the main results page shows a one-layer network, future development should aim to include a most extensive viewer that shows multiple levels. Additionally, as the predictor is further developed, the data included should be updated on a more regular basis to reflect the current state of PIPs and ensure that it remains an up-to-date tool for outside researchers to make use of.

## 6.4 Conclusions

1) The Nucleolar web server database has been updated to reflect the current state of the Interactome database and include the most recent evidence, protein and predicted interactions data.

2) The PIPs web server framework has been rewritten from its previous form of JSP to implement a PHP/MySQL/HTML framework.

3) The entire site has been restyled with a clean, modern and intuitive feel through HTML5 and CSS3.

4) Site functionality and, where applicable, previous optimisation has been preserved.

# Chapter 7

## Conclusions and Future Directions

## Preface

This chapter summarises the results contained in each chapter of this thesis. Additionally, suggestions are offered for future development for each aspect and the PIPs and PIP'NN predictors.

# 7.1 Further Developments to PIPs

This thesis has described the developments undertaken to advance the PIPs predictor (Chapter 2), the introduction of PIP'NN, a predictor built off of PIPs with a neural network in place of the naive Bayesian framework (Chapter 3), a comparison of the PIPs and PIP'NN methods both against each other and against other current human protein-protein interaction predictors (Chapter 4), practical applications of both predictors (Chapter 5), further extension of the predictor into cross-species prediction (Chapter 6) and the redesign of the PIPs web server (Chapter 7).

## 7.1.1 The PIPs Framework

First, the update of data considered the Orthology and Combined modules and the positive dataset has brought PIPs to a more up-to-date status (Chapters 2.2.1-2.2.6). However, as data is only ever current for a short period of time, maintenance of PIPs will require continued attention to monitor the status of this data. As PIPs, like other evidence-based prediction methods, will only ever be as strong as the data provided to it, this maintenance, along with retraining the predictor on a reasonably frequent basis, is important to keeping the tool as strong as possible. However, as the entire process of retraining, testing and predicting with PIPs is non-trivial and time-extensive, a balance between maintenance and further development should be established for maximum effectiveness.

The addition of the TransMCL module and subsequent development of the ECOZ predictor has had mixed success (Chapter 2.3). Initially, it was hoped that combining the Transitive and Cluster modules would have a cumulative effect of both increasing

the number and accuracy of predictions. However, while the EOCZ predictor both predicted nearly all of the interactions predicted by the EOCT and EOCM predictors independently and identified a large number of distinct interactions, this increase of coverage came with some compromise to the overall prediction accuracy. This decrease in accuracy was likely due to the Cluster module component, which on its own performs worse than the Transitive module on multiple blind tests. Ultimately, it was decided that the EOCT predictor, the original predictor from PIPs v. 1.0, would remain as the primary method for the Bayesian version of PIPs and the EOCM and EOCZ would be offered as additional options.

### 7.1.2 PIP'NN

The neural network version of PIPs, PIP'NN, has shown initial promise of becoming an alternate method of prediction (Chapter 3.3). However, determining a strict, final cut-off threshold for the PIP'NN prediction set proved difficult. As a result, a cut-off output score of 0.5 has been suggested as the initial cut-off, though users should be advised that the higher the output score, the more likely the interaction (Chapter 3.3.2.1). Compared to PIPs, PIP'NN was able to identify more known positive and negative interactions correctly and consistently in multiple blind test sets of varying sizes and compositions (Chapter 4.3).

There are two aspects of PIP'NN that will warrant attention for further development. First, with one output node giving one final output score that is a value between 0.0 and 1.0, there is no method for determining how confident a given prediction is. While it could be assumed that predictions with the highest output scores are strong predictions,

more difficultly arises with pairs scored in the mid-range. Therefore, a strict cut-off for what is a prediction and what is not is likely to over-predict and include a large number of false positives. As a result, like in all prediction frameworks, care should be taken in assessing predictions for practical use to ensure that pairs of interest are probable interactors. One option for further development would be to design a network with two output nodes, where a pair is assigned a score for 'No Predicted Interaction' and a score for 'Predicted Interaction'. In this case, the difference between the two scores could be assessed, such that the larger the difference, the stronger the prediction.

Second, while attempts were made to include the network analysis component of the original version of PIPs into PIP'NN, both taking the likelihood ratios calculated by Bayesian PIPs and incorporating the Transitive module as a second stage of analysis failed to improve performance of the method on blind tests (Chapters 3.3.2-3.3.3). Although it was at first thought that the 0.5 cut-off threshold originally selected for construction of the initial predicted interaction network to be analysed by the transitive component was too low, increasing the cut-off to 0.7 had a detrimental effect on the performance of the two-stage neural network predictor. With further consideration, it was determined that the network assembled with the 0.5 cut-off included a comparable number of pairs to the network in the Bayesian version of PIPs considered by the Transitive module. However, the inclusion of the transitive component with the two-stage network did not significantly increase the predictive capability of the neural network, and it was not included. Overall, it is likely that the evidence considered by the one-stage network is, on its own, enough for prediction without an extra step of analysis. As the network analysis component is an aspect of PIPs unique to other protein-protein interaction predictors and has been shown to have a positive effect on

predicting interactions, more effort could be taken to identify an alternate approach to allow its incorporation.

PIP'NN consistently performed more accurately than PIPs across a range of blind tests and analyses. However, as PIPs did not perform poorly but only worse comparably to PIP'NN, the predictions resulting from PIPs should not be discounted entirely, as there are distinct interactions predicted by the two methods. Therefore, the most effective way to implement PIP'NN is suggested to be in conjunction with the EOCT, EOCM and/or EOCZ PIPs predictors to identify both overlapping and unique potential interactions.

Finally, like in PIPs on its own, the predictions made by PIP'NN will only ever be as strong as the available evidence; therefore, further maintenance and development of the predictor should follow a similar protocol of data updates and retraining as PIPs.

Overall, the development of PIP'NN has shown that a neural network framework is able to successfully handle the prediction of protein-protein interactions. Compared to PIPs and other protein-protein interaction prediction methods currently available, PIP'NN consistently performs above average. This success is likely due to the neural network strategy of predicting based on patterns of evidence rather than individual pieces independently and has suggested a new direction for the field of protein-protein interaction to explore in the future.

### 7.1.3 Practical Application of PIPs and PIP'NN

The two collaborations with the Rouse (Chapter 5.3.1) and Lamond (Chapter 5.3.2) labs at the College of Life Sciences at the University of Dundee have highlighted examples of how PIPs and PIP'NN could be applied practically. While the predictions returned for the DNA repair proteins of interest to the Rouse lab were not able to be confirmed experimentally, the protocol established for identifying interactions serves as a framework that other labs could follow to make use of the predictors. Of particular note, it was necessary with this investigation to lower the cut-off threshold in order to return interactions for the majority of proteins of interest. This adjustment again brings to light the important point that the effectiveness of PIPs and PIP'NN, as evidence-based methods, is directly dependent on how much data is available. While specific proteins of interest may have sparse evidence available, homologues or similar proteins that have been more studied may also be worth considering for interaction prediction. Likewise, predictions for other proteins within the system, even if they are known not to interact with the target protein, may provide suggestions for potential interactors.

Incorporating PIPs and PIP'NN as an additional filtering step in the SILAC complex detection protocol has shown promising results for identifying protein complexes with low M/L and H/L SILAC ratios that may be due to genuine interactions with the target protein and not due to background noise or non-specific binding. Of the complexes in the low-ratio dataset examined for CUL4B, several predicted by PIPs and PIP'NN appear to be reasonable interactions based on the their functional and GO term annotations. In the future, PIPs and PIP'NN could easily be included as a standard step in the SILAC procedure to suggest the most reasonable interactions to follow up on with experimental confirmation.

### 7.1.4 The PIPs Web Server

The newly redesigned PIPs web server has brought the publicly accessible version of PIPs in line with the current in-house version of the predictor. In re-engineering the site, care has been taken in maintaining as much from previous development as possible. The new version of the PIPs web server offers users the ability to search for predictions from the EOCT, EOCM or EOCZ methods for proteins of interest. Results are returned in a clear manner for easy browsing.

In the future, the website should be maintained to reflect the most current, stable version of the predictor.

## 7.2 Future Directions for PIPs and PIP'NN

While the comparison with other currently available human protein-protein interaction prediction methods (Chapter 4.3.5) showed that PIPs and PIP'NN were able to predict comparable numbers of interactions within the selected set, PrePPI (Zhang *et al.*, 2012), a new method that considers evidence in a naive Bayesian framework similar to PIPs but also includes a structure module, far outperformed all other methods. Although sequence and structural properties have been considered both in PIPs and in other prediction methods, PrePPI attempts to circumvent the issue of not all proteins having solved structures by assigning each protein with a homologous model structure. The much higher number of predictions from the comparison test set by PrePPI indicates that this method of structure inclusion is a strong indicator of interaction.

Previously, the Barton Group has developed SNAPPI-DB, a database of atom-level domain-domain interactions based on structural data derived from the MSD database from the European Bioinformatics Institute (EBI) (Jefferson *et al.*, 2007). With this resource in place, PIPs could be extended to include a 'Structure' module based on these domain-domain interfaces. Although the full structures for main proteins may not yet be solved, in terms of interaction, the interfaces between the proteins are most crucial. Therefore, the domain interaction information provided by SNAPPI-DB should offer information to cover the majority of protein pairs. While domains frequently seen in interacting proteins are already included in the Combined module, they are only considered on a superficial level. However, before including a separate Structure module, the independence between the domain information included in SNAPPI-DB and in the Combined module would have to be considered. Depending on the results of this test, incorporation of the evidence could then take two methods: 1) by either replacing the domain co-occurrence portion of the Combined module or 2) as a separate module alongside of the Expression, Orthology and Combined modules.

The next major direction for PIPs and PIP'NN to take is to move further into the network resulting from predictions to construct a proposed map of the human interactome. Such a development should include not only interactions predicted by PIPs and PIP'NN, but should also build upon known, annotated interactions. While the STRING database currently acts as the main protein-protein interaction resource by providing this information, the PIPs and PIP'NN prediction component would offer an additional set of interactions to include. In particular, a strong method of dynamic visualisation for the predicted network will be crucial. While lists of predicted interactors can be sufficient on a small scale, actively seeing how the interactions fit

within the broader context of the network and being able to manipulate that network will provide valuable insight, both into how likely they are to be interactions or in identifying new associations between proteins that might not otherwise be considered.

Overall, PIPs and PIP'NN have reached a stable level where they can be implemented practically. While continued maintenance and further development will be necessary to allow the methods to have the maximum impact, this work should be completed in conjunction with engaging in active, experimental investigations. Although there remain novel future directions for PIPs and PIP'NN to explore, this thesis has continued to lay the foundations for this further progress and for making an effective predictor of human protein-protein interactions.

# Bibliography

Abu-Farha, M., Elisma, F. & Figeys, D. (2008) Identification of protein-protein interactions by mass spectrometry coupled techniques. *Advances in biochemical engineering/biotechnology*. [Online] 11067–80. Available from: doi:10.1007/10_2007_091.

Alexeyenko, A. & Sonnhammer, E.L.L. (2009) Global networks of functional coupling in eukaryotes from comprehensive data integration. *Genome research*. [Online] 19 (6), 1107–1116. Available from: doi:10.1101/gr.087528.108.

Alexeyenko, A., Schmitt, T., Tjärnberg, A., Guala, D., Frings, O. & Sonnhammer, E.L.L. (2012) Comparative interactomics with Funcoup 2.0. *Nucleic acids research*. [Online] 40 (Database issue), 821–828. Available from: doi:10.1093/nar/gkr1062.

Andreeva, A., Howorth, D., Chandonia, J.-M., Brenner, S.E., Hubbard, T.J.P., Chothia, C. & Murzin, A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic acids research*. [Online] 36 (Database issue), D419–D425. Available from: doi:10.1093/nar/gkm993.

Ashburner, M. & Lewis, S. (2002) On ontologies for biologists: the Gene Ontology--untangling the web. *Novartis Foundation symposium*. 24766–80; discussion80–discussion83, 84–90, 244–252.

Azam, F. (2000) Biologically inspired modular neural networks. *PhD Thesis Virginia Polytechnic Institute and State University*.

Bader, G.D., Betel, D. & Hogue, C.W.V. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic acids research*. 31 (1), 248–250.

Basheer, I.A. & Hajmeer, M. (2000) Artificial neural networks: fundamentals, computing, design, and application. *Journal of microbiological methods*. 43 (1), 3–31.

Ben-Hur, A. & Noble, W.S. (2006) Choosing negative examples for the prediction of protein-protein interactions. *BMC bioinformatics*. [Online] 7 Suppl 1S2. Available from: doi:10.1186/1471-2105-7-S1-S2.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. & Bourne, P.E. (2000) The Protein Data Bank. *Nucleic acids research*. 28 (1), 235–242.

Bhardwaj, N. & Lu, H. (2005) Correlation between gene expression profiles and protein-protein interactions within and across genomes. *Bioinformatics (Oxford, England)*. [Online] 21 (11), 2730–2738. Available from: doi:10.1093/bioinformatics/bti398.

Bock, J.R. & Gough, D.A. (2001) Predicting protein--protein interactions from primary structure. *Bioinformatics (Oxford, England)*. 17 (5), 455–460.

Bodén, M. & Hawkins, J. (2005) Prediction of subcellular localization using sequence-biased recurrent networks. *Bioinformatics (Oxford, England)*. [Online] 21 (10), 2279–2286. Available from: doi:10.1093/bioinformatics/bti372.

Boisvert, F.-M., Ahmad, Y., Gierliński, M., Charrière, F., Lamont, D., Scott, M., Barton, G. & Lamond, A.I. (2012) A quantitative spatial proteomics analysis of proteome turnover in human cells. *Molecular & cellular proteomics : MCP*. [Online] 11 (3), M111.011429. Available from: doi:10.1074/mcp.M111.011429.

Boulon, S., Ahmad, Y., Trinkle-Mulcahy, L., Verheggen, C., Cobley, A., Gregor, P., Bertrand, E., Whitehorn, M. & Lamond, A.I. (2010) Establishment of a protein frequency library and its application in the reliable identification of specific protein interaction partners. *Molecular & cellular proteomics : MCP*. [Online] 9 (5), 861–879. Available from: doi:10.1074/mcp.M900517-MCP200.

Breiman, L. (2001) Machine Learning, Volume 45, Number 1 - SpringerLink. *Machine Learning*. [Online] 45 (1), 5–32. Available from: doi:10.1023/A:1010933404324.

Brohée, S. & van Helden, J. (2006) Evaluation of clustering algorithms for protein-protein interaction networks. *BMC bioinformatics*. [Online] 7488. Available from: doi:10.1186/1471-2105-7-488.

Brown, K.R. & Jurisica, I. (2005) Online predicted human interaction database. *Bioinformatics (Oxford, England)*. [Online] 21 (9), 2076–2082. Available from: doi:10.1093/bioinformatics/bti273.

Brown, K.R. & Jurisica, I. (2007) Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biology*. [Online] 8 (5), R95. Available from: doi:10.1186/gb-2007-8-5-r95.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T.L. (2009) BLAST+: architecture and applications. *BMC bioinformatics*. [Online] 10421. Available from: doi:10.1186/1471-2105-10-421.

Chen, J.Y., Mamidipalli, S. & Huan, T. (2009) HAPPI: an online database of comprehensive human annotated and predicted protein interactions. *BMC genomics*. [Online] 10 Suppl 1S16. Available from: doi:10.1186/1471-2164-10-S1-S16.

Chinnasamy, A., Mittal, A. & Sung, W.-K. (2006) Probabilistic prediction of protein-protein interactions from the protein sequences. *Computers in biology and medicine*. [Online] 36 (10), 1143–1154. Available from: doi:10.1016/j.compbiomed.2005.09.005.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. & de Hoon, M.J.L. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)*. [Online] 25 (11), 1422–1423. Available from: doi:10.1093/bioinformatics/btp163.

Cole, C., Barber, J.D. & Barton, G.J. (2008) *The Jpred 3 secondary structure prediction server*. [Online] 36 (Web Server issue), W197–W201. Available from:

doi:10.1093/nar/gkn238.

Conte, Lo, L., Ailey, B., Hubbard, T.J., Brenner, S.E., Murzin, A.G. & Chothia, C. (2000) SCOP: a structural classification of proteins database. *Nucleic acids research*. 28 (1), 257–259.

Couto, F., Silva, M. & Coutinho, P. (2007) Measuring semantic similarity between Gene Ontology terms. *Data & Knowledge Engineering*. [Online] 61137–152. Available from: doi:doi:10.1016/j.datak.2006.05.003.

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., Mahajan, S., May, B., et al. (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research*. [Online] 39 (Database issue), D691–D697. Available from: doi:10.1093/nar/gkq1018.

Cybulski, K.E. & Howlett, N.G. (2011) FANCP/SLX4: a Swiss army knife of DNA interstrand crosslink repair. *Cell cycle (Georgetown, Tex)*. 10 (11), 1757–1763.

Dandekar, T., Snel, B., Huynen, M. & Bork, P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in biochemical sciences*. 23 (9), 324–328.

Deane, C.M. (2002) Protein Interactions: Two Methods for Assessment of the Reliability of High Throughput Observations. *Molecular & Cellular Proteomics*. [Online] 1 (5), 349–356. Available from: doi:10.1074/mcp.M100037-MCP200.

Duro, E., Lundin, C., Ask, K., Sanchez-Pulido, L., MacArtney, T.J., Toth, R., Ponting, C.P., Groth, A., Helleday, T. & Rouse, J. (2010) Identification of the MMS22L-TONSL complex that promotes homologous recombination. *Molecular Cell*. [Online] 40 (4), 632–644. Available from: doi:10.1016/j.molcel.2010.10.023.

Eliceiri, G.L. (1999) Small nucleolar RNAs. *Cellular and molecular life sciences : CMLS*. 56 (1-2), 22–31.

Emanuelsson, O., Nielsen, H., Brunak, S. & Heijne, von, G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of molecular biology*. [Online] 300 (4), 1005–1016. Available from: doi:10.1006/jmbi.2000.3903.

ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., Kuehn, M.S., Taylor, C.M., Neph, S., et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*. [Online] 447 (7146), 799–816. Available from: doi:10.1038/nature05874.

Enright, A.J. & Ouzounis, C.A. (2001) Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. *Genome Biology*. 2 (9), RESEARCH0034.

Enright, A.J., Iliopoulos, I., Kyrpides, N.C. & Ouzounis, C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*. [Online] 402 (6757), 86–90. Available from: doi:10.1038/47056.

Enright, A.J., Van Dongen, S. & Ouzounis, C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*. 30 (7), 1575–1584.

Ewing, R.M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., McBroom-Cerajewski, L., Robinson, M.D., O'Connor, L., Li, M., Taylor, R., Dharsee, M., Ho, Y., Heilbut, A., et al. (2007) Large-scale mapping of human protein-protein interactions by mass spectrometry. *Molecular systems biology*. [Online] 389. Available from: doi:10.1038/msb4100134.

Fariselli, P., Pazos, F., Valencia, A. & Casadio, R. (2002) Prediction of protein--protein interaction sites in heterocomplexes with neural networks. *European journal of biochemistry / FEBS*. 269 (5), 1356–1361.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Gordon, L., Hendrix, M., Hourlier, T., et al. (2012) Ensembl 2012. *Nucleic acids research*. [Online] 40 (Database issue), D84–D90. Available from: doi:10.1093/nar/gkr991.

Frank, J. (2002) Single-particle imaging of macromolecules by cryo-electron microscopy. *Annual review of biophysics and biomolecular structure*. [Online] 31303–319. Available from: doi:10.1146/annurev.biophys.31.082901.134202.

Gandhi, T.K.B., Zhong, J., Mathivanan, S., Karthick, L., Chandrika, K.N., Mohan, S.S., Sharma, S., Pinkert, S., Nagaraju, S., Periaswamy, B., Mishra, G., Nandakumar, K., Shen, B., Deshpande, N., et al. (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature genetics*. [Online] 38 (3), 285. Available from: doi:doi:10.1038/ng1747.

Garcia-Garcia, J., Schleker, S., Klein-Seetharaman, J. & Oliva, B. (2012) BIPS: BIANA Interolog Prediction Server. A tool for protein-protein interaction inference. *Nucleic acids research*. [Online] 40 (Web Server issue), W147–W151. Available from: doi:10.1093/nar/gks553.

Gardner, P.P., Bateman, A. & Poole, A.M. (2010) SnoPatrol: how many snoRNA genes are there? *Journal of biology*. [Online] 9 (1), 4. Available from: doi:10.1186/jbiol211.

Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.-M., Cruciat, C.-M., Remor, M., Höfert, C., Schelder, M., Brajenovic, M., et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*. [Online] 415 (6868), 141–147. Available from: doi:10.1038/415141a.

Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., et al. (2003) A protein interaction map of Drosophila melanogaster. *Science (New York, NY)*. [Online] 302 (5651), 1727–1736. Available from:

doi:10.1126/science.1090289.

Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J. & Lopez, R. (2010) A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic acids research*. [Online] 38 (Web Server issue), W695–W699. Available from: doi:10.1093/nar/gkq313.

Gurney, K., 2003. *Neural Networks.* Digital ed. London: Taylor and Francis.

Hahn, A., Rahnenführer, J., Talwar, P. & Lengauer, T. (2005) Confirmation of human protein interaction data by human expression data. *BMC bioinformatics*. [Online] 6112. Available from: doi:10.1186/1471-2105-6-112.

Hamilton, N., Burrage, K., Ragan, M.A. & Huber, T. (2004) Protein contact prediction using patterns of correlation. *Proteins*. [Online] 56 (4), 679–684. Available from: doi:10.1002/prot.20160.

Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G.M., Blake, J.A., Bult, C., et al. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic acids research*. [Online] 32 (Database issue), D258–D261. Available from: doi:10.1093/nar/gkh036.

Harrow, J.J., Frankish, A.A., Gonzalez, J.M.J., Tapanari, E.E., Diekhans, M.M., Kokocinski, F.F., Aken, B.L.B., Barrell, D.D., Zadissa, A.A., Searle, S.S., Barnes, I.I., Bignell, A.A., Boychenko, V.V., Hunt, T.T., et al. (2012) GENCODE: The reference human genome annotation for The ENCODE Project. *Genes & development*. [Online] 22 (9), 1760–1774. Available from: doi:10.1101/gr.135350.111.

Hecht-Nielsen, R. (1987) Counterpropagation networks. *Applied optics*. 26 (23), 4979–4983.

Hecht-Nielsen, R. (1990) *Theory of the backpropagation neural network.* I–593–I–605.

Higa, L.A., Wu, M., Ye, T., Kobayashi, R., Sun, H. & Zhang, H. (2006) CUL4-DDB1 ubiquitin ligase interacts with multiple WD40-repeat proteins and regulates histone methylation. *Nature cell biology*. [Online] 8 (11), 1277–1283. Available from: doi:10.1038/ncb1490.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.-L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., et al. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*. [Online] 415 (6868), 180–183. Available from: doi:10.1038/415180a.

Hue, M., Riffle, M., Vert, J.-P. & Noble, W.S. (2010) Large-scale prediction of protein-protein interactions from structures. *BMC bioinformatics*. [Online] 11144. Available from: doi:10.1186/1471-2105-11-144.

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T.K., Bateman, A., Bernard,

T., Binns, D., Bork, P., Burge, S., de Castro, E., Coggill, P., Corbett, M., Das, U., et al. (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic acids research*. [Online] 40 (Database issue), D306–D312. Available from: doi:10.1093/nar/gkr948.

Imanishi, M., Imamura, C., Higashi, C., Yan, W., Negi, S., Futaki, S. & Sugiura, Y. (2010) Zinc finger-zinc finger interaction between the transcription factors, GATA-1 and Sp1. *Biochemical and biophysical research communications*. [Online] 400 (4), 625–630. Available from: doi:10.1016/j.bbrc.2010.08.116.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F. & Gerstein, M. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science (New York, NY)*. [Online] 302 (5644), 449–453. Available from: doi:10.1126/science.1087361.

Jefferson, E.R., Walsh, T.P., Roberts, T.J. & Barton, G.J. (2007) SNAPPI-DB: a database and API of Structures, iNterfaces and Alignments for Protein-Protein Interactions. *Nucleic acids research*. [Online] 35 (Database issue), D580–D589. Available from: doi:10.1093/nar/gkl836.

Jiang, J. & Conrath, D. (1997) Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings of Interational Conference Research on Computational Linguistics (ROCLING X)*.

Kelley, R. & Ideker, T. (2005) Systematic interpretation of genetic interactions using protein networks. *Computational Biology*. 23 (5), 561–566.

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., Jandrasits, C., Jimenez, R.C., Khadake, J., Mahadevan, U., et al. (2012) The IntAct molecular interaction database in 2012. *Nucleic acids research*. [Online] 40 (Database issue), D841–D846. Available from: doi:10.1093/nar/gkr1088.

Kersey, P.J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E. & Apweiler, R. (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics*. [Online] 4 (7), 1985–1988. Available from: doi:10.1002/pmic.200300721.

Kiemer, L. & Cesareni, G. (2007) Comparative interactomics: comparing apples and pears? *Trends in biotechnology*. [Online] 25 (10), 448–454. Available from: doi:10.1016/j.tibtech.2007.08.002.

Kishore, S. & Stamm, S. (2006) Regulation of alternative splicing by snoRNAs. *Cold Spring Harbor symposia on quantitative biology*. [Online] 71329–334. Available from: doi:10.1101/sqb.2006.71.024.

Knisley, D. & Knisley, J. (2011) Predicting protein-protein interactions using graph invariants and a neural network. *Computational biology and chemistry*. [Online] 35 (2), 108–113. Available from: doi:10.1016/j.compbiolchem.2011.03.003.

Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller,

R., Dreher, K., Alexander, D.L., Garcia-Hernandez, M., Karthikeyan, A.S., Lee, C.H., Nelson, W.D., Ploetz, L., et al. (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic acids research*. [Online] 40 (Database issue), D1202–D1210. Available from: doi:10.1093/nar/gkr1090.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., et al. (2001) Initial sequencing and analysis of the human genome. *Nature*. [Online] 409 (6822), 860–921. Available from: doi:10.1038/35057062.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J. & Higgins, D.G. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics (Oxford, England)*. [Online] 23 (21), 2947–2948. Available from: doi:10.1093/bioinformatics/btm404.

Lee, J. & Zhou, P. (2007) DCAFs, the missing link of the CUL4-DDB1 ubiquitin ligase. *Molecular Cell*. [Online] 26 (6), 775–780. Available from: doi:10.1016/j.molcel.2007.06.001.

Lee, J. & Zhou, P. (2012) Pathogenic Role of the CRL4 Ubiquitin Ligase in Human Disease. *Frontiers in oncology*. [Online] 221. Available from: doi:10.3389/fonc.2012.00021.

Lehne, B. & Schlitt, T. (2009) Protein-protein interaction databases: keeping up with growing interactomes. *Human Genomics*. 3 (3), 291–297.

Lehner, B. & Fraser, A. (2004) A first-draft human protein-interaction map. *Genome Biology*. [Online] 5 (9), R63–R63. Available from: doi:10.1186/gb-2004-5-9-r63.

Li, T., Chen, X., Garbutt, K.C., Zhou, P. & Zheng, N. (2006) Structure of DDB1 in complex with a paramyxovirus V protein: viral hijack of a propeller cluster in ubiquitin ligase. *Cell*. [Online] 124 (1), 105–117. Available from: doi:10.1016/j.cell.2005.10.033.

Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A.P., Santonico, E., Castagnoli, L. & Cesareni, G. (2012) MINT, the molecular interaction database: 2012 update. *Nucleic acids research*. [Online] 40 (Database issue), D857–D861. Available from: doi:10.1093/nar/gkr930.

Liu, J., Furukawa, M., Matsumoto, T. & Xiong, Y. (2002) NEDD8 modification of CUL1 dissociates p120(CAND1), an inhibitor of CUL1-SKP1 binding and SCF ligases. *Molecular Cell*. 10 (6), 1511–1518.

MacKay, C., Déclais, A.-C., Lundin, C., Agostinho, A., Deans, A.J., MacArtney, T.J., Hofmann, K., Gartner, A., West, S.C., Helleday, T., Lilley, D.M.J. & Rouse, J. (2010) Identification of KIAA1018/FAN1, a DNA repair nuclease recruited to DNA damage by monoubiquitinated FANCD2. *Cell*. [Online] 142 (1), 65–76. Available from: doi:10.1016/j.cell.2010.06.021.

Martin, S., Roe, D. & Faulon, J.-L. (2005) Predicting protein-protein interactions using

signature products. *Bioinformatics (Oxford, England)*. [Online] 21 (2), 218–226. Available from: doi:10.1093/bioinformatics/bth483.

Matthews, L.R. (2001) Identification of Potential Interaction Networks Using Sequence-Based Searches for Conserved Protein-Protein Interactions or 'Interologs'. *Genome research*. [Online] 11 (12), 2120–2126. Available from: doi:10.1101/gr.205301.

Mayer, U. (2008) Protein Information Crawler (PIC): extensive spidering of multiple protein information resources for large protein sets. *Proteomics*. [Online] 8 (1), 42–44. Available from: doi:10.1002/pmic.200700865.

McDowall, J. & Hunter, S. (2011) InterPro protein classification. *Methods in molecular biology (Clifton, NJ)*. [Online] 69437–47. Available from: doi:10.1007/978-1-60761-977-2_3.

McDowall, M.D. (2011) *Human Protein-Protein Interaction Prediction*. 1–264.

McDowall, M.D., Scott, M.S. & Barton, G.J. (2009) PIPs: human protein-protein interaction prediction database. *Nucleic acids research*. [Online] 37 (Database issue), D651–D656. Available from: doi:10.1093/nar/gkn870.

Mering, C.V., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A. & Bork, P. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic acids research*. [Online] 33 (Database issue), D433–D437. Available from: doi:10.1093/nar/gki005.

Mering, C.V., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. & Bork, P. (2002) Comparative assessment of large-scale data sets of

protein|[ndash]|protein interactions. *Nature*. [Online] 417 (6887), 399. Available from: doi:doi:10.1038/nature750.

Metz, C.E. (1978) Basic principles of ROC analysis. *Seminars in nuclear medicine*. 8 (4), 283–298.

Mewes, H.W., Ruepp, A., Theis, F., Rattei, T., Walter, M., Frishman, D., Suhre, K., Spannagl, M., Mayer, K.F.X., Stümpflen, V. & Antonov, A. (2011) MIPS: curated databases and comprehensive secondary data resources in 2010. *Nucleic acids research*. [Online] 39 (Database issue), D220–D224. Available from: doi:10.1093/nar/gkq1157.

Mooney, C., Wang, Y.-H. & Pollastri, G. (2011) SCLpred: protein subcellular localization prediction by N-to-1 neural networks. *Bioinformatics (Oxford, England)*. [Online] 27 (20), 2812–2819. Available from: doi:10.1093/bioinformatics/btr494.

Morgenstern, B. (2004) DIALIGN: multiple DNA and protein sequence alignment at BiBiServ. *Nucleic acids research*. [Online] 32 (Web Server issue), W33–W36. Available from: doi:10.1093/nar/gkh373.

Muñoz, I.M., Hain, K., Déclais, A.-C., Gardiner, M., Toh, G.W., Sanchez-Pulido, L., Heuckmann, J.M., Toth, R., Macartney, T., Eppink, B., Kanaar, R., Ponting, C.P., Lilley, D.M.J. & Rouse, J. (2009) Coordination of structure-specific nucleases by human SLX4/BTBD12 is required for DNA repair. *Molecular Cell*. [Online] 35 (1), 116–127. Available from: doi:10.1016/j.molcel.2009.06.020.

Nakagawa, T. & Xiong, Y. (2011) X-linked mental retardation gene CUL4B targets ubiquitylation of H3K4 methyltransferase component WDR5 and regulates neuronal gene expression. *Molecular Cell*. [Online] 43 (3), 381–391. Available from: doi:10.1016/j.molcel.2011.05.033.

Nicol, J.W., Helt, G.A., Blanchard, S.G., Raja, A. & Loraine, A.E. (2009) The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics (Oxford, England)*. [Online] 25 (20), 2730–2731. Available from: doi:10.1093/bioinformatics/btp472.

O'Brien, K.P., Remm, M. & Sonnhammer, E.L.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic acids research*. [Online] 33 (Database issue), D476–D480. Available from: doi:10.1093/nar/gki107.

O'Connell, M.R., Gamsjaeger, R. & Mackay, J.P. (2009) The structural analysis of protein-protein interactions by NMR spectroscopy. *Proteomics*. [Online] 9 (23), 5224–5232. Available from: doi:10.1002/pmic.200900303.

Ofran, Y. & Rost, B. (2003) Predicted protein-protein interaction sites from local sequence information. *FEBS letters*. 544 (1-3), 236–239.

Ooi, S.L., Pan, X., Peyser, B.D., Ye, P., Meluh, P.B., Yuan, D.S., Irizarry, R.A., Bader, J.S., Spencer, F.A. & Boeke, J.D. (2006) Global synthetic-lethality analysis and yeast functional profiling. *Trends in genetics : TIG*. [Online] 22 (1), 56–63. Available from: doi:10.1016/j.tig.2005.11.003.

Osaka, F., Kawasaki, H., Aida, N., Saeki, M., Chiba, T., Kawashima, S., Tanaka, K. & Kato, S. (1998) A new NEDD8-ligating system for cullin-4A. *Genes & development*. 12 (15), 2263–2268.

Ostlund, G., Schmitt, T., Forslund, K., Köstler, T., Messina, D.N., Roopra, S., Frings, O. & Sonnhammer, E.L.L. (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic acids research*. [Online] 38 (Database issue), D196–D203. Available from: doi:10.1093/nar/gkp931.

Pan, Z.-Q., Kentsis, A., Dias, D.C., Yamoah, K. & Wu, K. (2004) Nedd8 on cullin: building an expressway to protein destruction. *Oncogene*. [Online] 23 (11), 1985–1997. Available from: doi:10.1038/sj.onc.1207414.

Pavlidis, P., Weston, J., Cai, J. & Noble, W.S. (2002) Learning gene functional classifications from multiple data types. *Journal of computational biology : a journal of computational molecular cell biology*. [Online] 9 (2), 401–411. Available from: doi:10.1089/10665270252935539.

Pazos, F. & Valencia, A. (2008) Protein co-evolution, co-adaptation and interactions.

*The EMBO journal*. [Online] 27 (20), 2648–2655. Available from: doi:10.1038/emboj.2008.189.

Peri, S., Navarro, J.D., Amanchy, R., Kristiansen, T.Z., Jonnalagadda, C.K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T.K.B., Gronborg, M., Ibarrola, N., Deshpande, N., Shanker, K., Shivashankar, H.N., et al. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome research*. [Online] 13 (10), 2363–2371. Available from: doi:10.1101/gr.1680803.

Pitre, S., Alamgir, M., Green, J.R., Dumontier, M., Dehne, F. & Golshani, A. (2008) Computational methods for predicting protein-protein interactions. *Advances in biochemical engineering/biotechnology*. [Online] 110247–267. Available from: doi:10.1007/10_2007_089.

Prasad, T.S.K. (n.d.) *Human protein reference database—2009 update*.

Prieto, C. & Las Rivas, De, J. (2006) APID: Agile Protein Interaction DataAnalyzer. *Nucleic acids research*. [Online] 34 (Web Server issue), W298–W302. Available from: doi:10.1093/nar/gkl128.

Pruitt, K.D., Tatusova, T., Brown, G.R. & Maglott, D.R. (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic acids research*. [Online] 40 (Database issue), D130–D135. Available from: doi:10.1093/nar/gkr1079.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E.L.L., Eddy, S.R., et al. (2012) The Pfam protein families database. *Nucleic acids research*. [Online] 40 (Database issue), D290–D301. Available from: doi:10.1093/nar/gkr1065.

Qi, Y., Bar-Joseph, Z. & Klein-Seetharaman, J. (2006) Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins*. [Online] 63 (3), 490–500. Available from: doi:10.1002/prot.20865.

Qi, Y., Klein-Seetharaman, J. & Bar-Joseph, Z. (2007) A mixture of feature experts approach for protein-protein interaction prediction. *BMC bioinformatics*. [Online] 8 Suppl 10S6. Available from: doi:10.1186/1471-2105-8-S10-S6.

Qi, Y., Klein-Seetharaman, J. & Bar-Joseph, Z. (2005) Random forest similarity for protein-protein interaction prediction from multiple sources. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*. 531–542.

Rajagopala, S.V., Sikorski, P., Caufield, J.H., Tovchigrechko, A. & Uetz, P. (2012) Studying protein complexes by the yeast two-hybrid system. *Methods (San Diego, Calif)*. [Online] Available from: doi:10.1016/j.ymeth.2012.07.015.

Reimand, J., Hui, S., Jain, S., Law, B. & Bader, G.D. (2012) Domain-mediated protein interaction prediction: From genome to network. *FEBS letters*. [Online] 586 (17),

2751–2763. Available from: doi:10.1016/j.febslet.2012.04.027.

Remm, M., Storm, C.E. & Sonnhammer, E.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of molecular biology*. [Online] 314 (5), 1041–1052. Available from: doi:10.1006/jmbi.2000.5197.

Rhodes, D.R., Tomlins, S.A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A. & Chinnaiyan, A.M. (2005) Probabilistic model of the human protein-protein interaction network. *Nature biotechnology*. [Online] 23 (8), 951–959. Available from: doi:10.1038/nbt1103.

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M. & Séraphin, B. (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nature biotechnology*. [Online] 17 (10), 1030–1032. Available from: doi:10.1038/13732.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C. & Müller, M. (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics*. [Online] 1277. Available from: doi:10.1186/1471-2105-12-77.

ROSENBLATT, F. (1958) The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*. 65 (6), 386–408.

Roslan, R., Othman, R.M., Shah, Z.A., Kasim, S., Asmuni, H., Taliba, J., Hassan, R. & Zakaria, Z. (2010) Utilizing shared interacting domain patterns and Gene Ontology information to improve protein-protein interaction prediction. *Computers in biology and medicine*. [Online] 40 (6), 555–564. Available from: doi:10.1016/j.compbiomed.2010.03.009.

Rouse, J. (2009) Control of genome stability by SLX protein complexes. *Biochemical Society transactions*. [Online]. 37 (Pt 3) 495–510. Available from: doi:10.1042/BST0370495.

Russell, R.B., Alber, F., Aloy, P., Davis, F.P., Korkin, D., Pichaud, M., Topf, M. & Sali, A. (2004) A structural perspective on protein-protein interactions. *Current opinion in structural biology*. [Online] 14 (3), 313–324. Available from: doi:10.1016/j.sbi.2004.04.006.

Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., Nativ, N., Bahir, I., Doniger, T., Krug, H., Sirota-Madi, A., Olender, T., Golan, Y., Stelzer, G., et al. (2010) GeneCards Version 3: the human gene integrator. *Database : the journal of biological databases and curation*. [Online] 2010baq020. Available from: doi:10.1093/database/baq020.

Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U. & Eisenberg, D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic acids research*. [Online] 32 (Database issue), D449–D451. Available from: doi:10.1093/nar/gkh086.

Sarikas, A., Hartmann, T. & Pan, Z.-Q. (2011) The cullin protein family. *Genome Biology*. [Online] 12 (4), 220. Available from: doi:10.1186/gb-2011-12-4-220.

Schmitt, T., Messina, D.N., Schreiber, F. & Sonnhammer, E.L.L. (2011) Letter to the editor: SeqXML and OrthoXML: standards for sequence and orthology information. *Briefings in bioinformatics*. [Online] 12 (5), 485–488. Available from: doi:10.1093/bib/bbr025.

Scott, M.S. & Barton, G.J. (2007) Probabilistic prediction and ranking of human protein-protein interactions. *BMC bioinformatics*. [Online] 8239. Available from: doi:10.1186/1471-2105-8-239.

Scott, M.S., Troshin, P.V. & Barton, G.J. (2011) NoD: a Nucleolar localization sequence detector for eukaryotic and viral proteins. *BMC bioinformatics*. [Online] 12317. Available from: doi:10.1186/1471-2105-12-317.

Shoemaker, B.A. & Panchenko, A.R. (2007) Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS computational biology*. [Online] 3 (4), e43. Available from: doi:10.1371/journal.pcbi.0030043.

Smialowski, P., Pagel, P., Wong, P., Brauner, B., Dunger, I., Fobo, G., Frishman, G., Montrone, C., Rattei, T., Frishman, D. & Ruepp, A. (2010) The Negatome database: a reference set of non-interacting protein pairs. *Nucleic acids research*. [Online] 38 (Database issue), D540–D544. Available from: doi:10.1093/nar/gkp1026.

Smyth, M.S. & Martin, J.H. (2000) x ray crystallography. *Molecular pathology : MP*. 53 (1), 8–14.

Snel, B., van Noort, V. & Huynen, M.A. (2004) Gene co-regulation is highly conserved in the evolution of eukaryotes and prokaryotes. *Nucleic acids research*. [Online] 32 (16), 4725–4731. Available from: doi:10.1093/nar/gkh815.

Spahn, C.M.T. & Penczek, P.A. (2009) Exploring conformational modes of macromolecular assemblies by multiparticle cryo-EM. *Current opinion in structural biology*. [Online] 19 (5), 623–631. Available from: doi:10.1016/j.sbi.2009.08.001.

Sprinzak, E. & Margalit, H. (2001) Correlated sequence-signatures as markers of protein-protein interaction. *Journal of molecular biology*. [Online] 311 (4), 681–692. Available from: doi:10.1006/jmbi.2001.4920.

Sprinzak, E., Sattath, S. & Margalit, H. (2003) How reliable are experimental protein-protein interaction data? *Journal of molecular biology*. 327 (5), 919–923.

Stark, C., Breitkreutz, B.-J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M.S., Nixon, J., Van Auken, K., Wang, X., Shi, X., Reguly, T., Rust, J.M., Winter, A., Dolinski, K., et al. (2011) The BioGRID Interaction Database: 2011 update. *Nucleic acids research*. [Online] 39 (Database issue), D698–D704. Available from: doi:10.1093/nar/gkq1116.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., et al. (2005) A human protein-protein interaction network:

a resource for annotating the proteome. *Cell*. [Online] 122 (6), 957–968. Available from: doi:10.1016/j.cell.2005.08.029.

Stumpf, M.P.H., Thorne, T., de Silva, E., Stewart, R., An, H.J., Lappe, M. & Wiuf, C. (2008) Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences of the United States of America*. [Online] 105 (19), 6959–6964. Available from: doi:10.1073/pnas.0708078105.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L.J. & Mering, C.V. (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*. [Online] 39 (Database issue), D561–D568. Available from: doi:10.1093/nar/gkq973.

Tarpey, P.S., Raymond, F.L., O'Meara, S., Edkins, S., Teague, J., Butler, A., Dicks, E., Stevens, C., Tofts, C., Avis, T., Barthorpe, S., Buck, G., Cole, J., Gray, K., et al. (2007) Mutations in CUL4B, which encodes a ubiquitin E3 ligase subunit, cause an X-linked mental retardation syndrome associated with aggressive outbursts, seizures, relative macrocephaly, central obesity, hypogonadism, pes cavus, and tremor. *American journal of human genetics*. [Online] 80 (2), 345–352. Available from: doi:10.1086/511134.

Teichmann, S.A. & Babu, M.M. (2002) Conservation of gene co-regulation in prokaryotes and eukaryotes. *Trends in biotechnology*. 20 (10), 407–410; discussion410.

Tirosh, I. & Barkai, N. (2005) Computational verification of protein-protein interactions by orthologous co-expression. *BMC bioinformatics*. [Online] 640. Available from: doi:10.1186/1471-2105-6-40.

Tong, A.H.Y. & Boone, C. (2006) Synthetic genetic array analysis in Saccharomyces cerevisiae. *Methods in molecular biology (Clifton, NJ)*. 313171–192.

Trinkle-Mulcahy, L., Boulon, S., Lam, Y.W., Urcia, R., Boisvert, F.-M., Vandermoere, F., Morrice, N.A., Swift, S., Rothbauer, U., Leonhardt, H. & Lamond, A. (2008) Identifying specific protein interaction partners using quantitative mass spectrometry and bead proteomes. *The Journal of cell biology*. [Online] 183 (2), 223–239. Available from: doi:10.1083/jcb.200805092.

Uetz, P. (2002) Two-hybrid arrays. *Current opinion in chemical biology*. 6 (1), 57–62.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., et al. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*. [Online] 403 (6770), 623–627. Available from: doi:10.1038/35001009.

UniProt Consortium (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic acids research*. [Online] 40 (Database issue), D71–D75. Available from: doi:10.1093/nar/gkr981.

Vinogradova, O. & Qin, J. (2012) NMR as a unique tool in assessment and complex determination of weak protein-protein interactions. *Topics in current chemistry*. [Online] 32635–45. Available from: doi:10.1007/128_2011_216.

Wagner, A. (2001) The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Molecular biology and evolution*. 18 (7), 1283–1292.

Walhout, A.J., Sordella, R., Lu, X., Hartley, J.L., Temple, G.F., Brasch, M.A., Thierry-Mieg, N. & Vidal, M. (2000) Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science (New York, NY)*. 287 (5450), 116–122.

Wiles, A.M., Doderer, M., Ruan, J., Gu, T.-T., Ravi, D., Blackman, B. & Bishop, A.J.R. (2010) Building and analyzing protein interactome networks by cross-species comparisons. *BMC systems biology*. [Online] 436. Available from: doi:10.1186/1752-0509-4-36.

Wilming, L.G., Gilbert, J.G.R., Howe, K., Trevanion, S., Hubbard, T. & Harrow, J.L. (2008) The vertebrate genome annotation (Vega) database. *Nucleic acids research*. [Online] 36 (Database issue), D753–D760. Available from: doi:10.1093/nar/gkm987.

Xia, J.-F., Zhao, X.-M. & Huang, D.-S. (2010) Predicting protein-protein interactions from protein sequences using meta predictor. *Amino acids*. [Online] 39 (5), 1595–1599. Available from: doi:10.1007/s00726-010-0588-1.

Xia, K., Dong, D. & Han, J.-D.J. (2006) IntNetDB v1.0: an integrated protein-protein interaction network database generated by a probabilistic model. *BMC bioinformatics*. [Online] 7508. Available from: doi:10.1186/1471-2105-7-508.

Xia, Y., Yu, H., Jansen, R., Seringhaus, M., Baxter, S., Greenbaum, D., Zhao, H. & Gerstein, M. (2004) Analyzing cellular biochemistry in terms of molecular networks. *Annual review of biochemistry*. [Online] 731051–1087. Available from: doi:10.1146/annurev.biochem.73.011303.073950.

Yamasaki, C., Murakami, K., Takeda, J.-I., Sato, Y., Noda, A., Sakate, R., Habara, T., Nakaoka, H., Todokoro, F., Matsuya, A., Imanishi, T. & Gojobori, T. (2010) H-InvDB in 2009: extended database and data mining resources for human genes and transcripts. *Nucleic acids research*. [Online] 38 (Database issue), D626–D632. Available from: doi:10.1093/nar/gkp1020.

Young, L., Jernigan, R.L. & Covell, D.G. (1994) A role for surface hydrophobicity in protein-protein recognition. *Protein science : a publication of the Protein Society*. [Online] 3 (5), 717–729. Available from: doi:10.1002/pro.5560030501.

Zaki, N., El-Hajj, W., Kamel, H.M. & Sibai, F. (2011) Strike: a protein-protein interaction classification approach. *Advances in experimental medicine and biology*. [Online] 696263–270. Available from: doi:10.1007/978-1-4419-7046-6_26.

Zell, A. (1995) *SNNS*.

Zhang, Q.C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C.A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T., Maniatis, T., Califano, A. & Honig, B. (2012) Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature*. [Online] Available from: doi:10.1038/nature11503.